LEARNING TO DEFER WHEN ROUTING CHANGES COMPETENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-to-defer systems assign inputs according to the current competence of the available models or experts. When models train only on the inputs they handle, routing also determines the training exposure that sustains future competence. We study how routing and training choices maintain useful capabilities in continually updated cascades. In a controlled language-model cascade, upgrading the fallback raises system accuracy from 90.0% to 93.0% while the smaller model's division accuracy falls from 64.4% to 16.2%. Training on the fallback's answers to deferred inputs preserves division accuracy at 65.2%, allowing competence to be maintained even when the fallback remains preferred. Further interventions show that restoring target examples at a fixed training budget brings performance close to the control, while parameter isolation prevents loss from other tasks' updates. An exposure-dependent model of cost-sensitive deferral gives sufficient conditions for a unique competence equilibrium in terms of router smoothness and variation in case-level thresholds, along with exposure requirements and recovery times. Image cascades show a complementary source of competence loss as examples from deferred classes leave the training window. These results support designing routing and training policies around both current performance and retained competence.

1 INTRODUCTION

Learning-to-defer (L2D) systems choose which model or expert should answer an input. Classical offline formulations optimize prediction and deferral for a fixed data distribution (Madras et al., 2018; Mozannar & Sontag, 2020). When a deployed model continues learning from the inputs it handles, assignment also determines training exposure and can change which decision-maker is preferred later.

For example, consider a model cascade that routes inputs to either a smaller model or a more accurate but more expensive fallback model. At a fixed cost of deferral, upgrading the fallback can improve overall accuracy while reducing the number of inputs handled by the smaller model. If the router favors the expert model for a particular task, the smaller model receives fewer training examples and has less opportunity to improve or maintain its accuracy. Reduced exposure may then widen the competence gap and make subsequent deferral more likely.

$$\text{routing} \longrightarrow \text{exposure} \longrightarrow \text{competence} \longrightarrow \text{routing}. \tag{1}$$

We connect cost-sensitive L2D with performative prediction to study competence maintenance in continually trained cascades (Mozannar & Sontag, 2020; Perdomo et al., 2020). In concurrent work, Huang & Vishnoi (2026) analyze coupled human skill and AI reliance. We focus on how selective training examples and parameter updates shape retained competence, and which routing and learning choices preserve it.

We distinguish which model handles an input from which models learn from its feedback. Training on deferred cases can preserve the smaller model's competence even when the fallback continues to handle those cases. We therefore evaluate retained task accuracy alongside system accuracy and fallback use. The aim is to preserve capabilities needed by the deployment, with their value weighed against the resources required to maintain them.

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.

Contributions. (i) **Identifying the training mechanism.** Controlled language-model interventions trace competence loss to reduced training on the affected task together with continued updates on other tasks. Restoring the task's examples at a fixed total training budget brings performance close to the control with the original fallback, while isolating its parameters prevents the measured loss from other tasks' updates.

(ii) **Conditions for persistence and recovery.** We couple cost-sensitive deferral with an exposure-dependent competence model to characterize when multiple stable competence levels arise. The analysis gives sufficient uniqueness conditions based on router smoothness and the distribution of case-level thresholds, along with the exposure and time needed for recovery and the window in which reversing a fallback upgrade restores the practiced state.

(iii) **Serving and training as separate decisions.** With feedback and learning dynamics that do not depend on serving assignment, we show that serving can remain greedy while training access is chosen separately, and quantify the serving cost this separation can avoid. Experiments at matched training doses compare independent training access, teacher feedback, replay, and exposure floors (minimum allocations to the smaller model), measuring retained competence separately from renewed assignment to the smaller model.

2 Related Work

Learning to defer. Learning-to-defer methods select between a predictor and an expert (Madras et al., 2018), with consistent surrogates for the joint prediction and deferral rule (Mozannar & Sontag, 2020). Recent formulations address changing expert pools (Duy et al., 2026) and nonstationary streams whose query decisions account for predictor updates (Montreuil et al., 2026b). Our focus is the retention and recovery of acquired capabilities under selective training exposure.

Online model cascades. Model routing and cascades reserve expensive models for selected inputs to reduce inference cost (Chen et al., 2023; Jitkrittum et al., 2023; Ong et al., 2024). Deferred queries can also support continued learning: Nie et al. (2024) update smaller models using large-model demonstrations, while Wu et al. (2025) retrieve strategies from past deferred queries without changing model parameters.

Continual learning and forgetting. Training on some tasks can impair performance on others (Yu et al., 2020; Kotha et al., 2024; Biderman et al., 2024), sometimes by disrupting task alignment (Zheng et al., 2025). Replay and task-specific parameters can help retain acquired capabilities (Rolnick et al., 2019; Houlsby et al., 2019). In our setting, routing determines which skills receive training examples. Our interventions restore target examples at a fixed total row count and control competing updates, identifying how selective exposure interacts with shared-parameter interference.

Performative prediction. Perdomo et al. (2020) formalize prediction when deployment changes the evaluation distribution. Extensions study stochastic optimization and stateful responses to past deployments (Mendler-Dünner et al., 2020; Brown et al., 2022). Related models link future training data to user retention (Hashimoto et al., 2018) and future rewards to prior selection (Metelli et al., 2022; Patil et al., 2023). Here, routing changes competence through selective training exposure and forgetting.

Skill and delegation under automation. Assistance can support continued learning (Bragg & Brunskill, 2020), while models of human–AI interaction examine how AI use changes skill, effort, and productivity (Aouad et al., 2026). Huang & Vishnoi (2026) analyze coupled human skill and AI reliance, including learning during delegation, and characterize distinct long-run skill regimes. Concurrently, we study competence maintenance in trained cascades. We derive conditions for equilibrium uniqueness and recovery under cost-sensitive routing and quantify the serving cost avoided by maintaining training access while the fallback answers.

3 ROUTING, EXPOSURE, AND COMPETENCE

3.1 SERVING AND LEARNING ARE TWO DECISIONS

Partition inputs into observable regions indexed by j, for example by task. At round t, let $a_{j,t}$ and $b_{j,t}$ be the student's and fallback's accuracy, and $\lambda_j \geq 0$ the penalty for a fallback call in accuracy units. The cost-sensitive greedy rule *keeps* a query with the student, rather than *deferring* it, when

$$\rho_{j,t} = \mathbf{1}\{\hat{a}_{j,t} \geq \theta_{j,t}\}, \qquad \theta_{j,t} = b_{j,t} - \lambda_j, \tag{2}$$

where $\hat{a}_{j,t}$ is the router's estimate of $a_{j,t}$ and $\theta_{j,t}$ is the routing threshold (the *bar*). We separate two controls: *assignment* $u_{j,t}$, the fraction of region-j queries the student answers, and *training access* or *exposure* $v_{j,t}$, the fraction that enter its update $w_{t+1} = \mathcal{A}(w_t, \mathcal{D}_t^{\text{train}})$. Handled-only learning imposes $v = u = \rho$. It is the regime in which feedback exists only for answers the student served, such as user corrections or verification of those answers; our experiments give every served answer its correct label, the most favourable case of that regime. The interventions we study change v with or without u: an *exposure floor* sets $u = v = \max\{\rho, \delta_j\}$ with *dose* δ_j; *training access* sets v while the fallback answers ($u = 0$); *teacher labels* do the same with the fallback's outputs as targets; *replay* supplies stored examples; and a *per-skill adapter* changes which parameters other regions' updates touch.

3.2 CONDITIONS FOR PERSISTENCE, PRESERVATION, AND RECOVERY

For one region, let competence $a \in [a_{\min}, a_{\max}]$ evolve as

$$\dot{a} = \kappa\, v\, (a_{\max} - a) - \mu\, (a - a_{\min}), \qquad \kappa, \mu > 0, \tag{3}$$

This scalar law is a tractable model of exposure-dependent competence. In the language-model setting, the decay term represents interference from other regions' updates through shared parameters; bounded-memory refitting supplies a different loss mechanism when target examples leave the training window. The law is not an exact derivation of adapter updates, and frozen or isolated target parameters remove the tested interference path. With constant exposure, competence converges to

$$a^*(v) = \frac{a_{\min} + \varphi v a_{\max}}{1 + \varphi v}, \qquad \varphi = \kappa/\mu. \tag{4}$$

Write $a_{\text{prac}} = a^*(1)$, the practiced level. Under handled-only greedy routing with exact estimates, $v(a) = \mathbf{1}\{a \geq \theta\}$.

Proposition 1 (Two self-consistent states)**.** *If $a_{\min} < \theta < a_{\text{prac}}$, both $a_{\min}$ and a_{prac} are attracting equilibria of the handled-only greedy dynamics; states below θ converge to $a_{\min}$ and states at or above it converge to a_{prac}. If $\theta > a_{\text{prac}}$, $a_{\min}$ is the only equilibrium.*

A fallback upgrade raises θ without touching a. If the new bar exceeds a_{prac}, the upper state disappears and entry is certain; if it lies below a_{prac} but above the current competence or its estimate, the skill can still be tipped into the lower of two coexisting states.

Corollary 2 (Rollback window)**.** *Let $a_{\min} < \theta_0 < a_{\text{prac}}$, let a skill sit at $a_0 \geq \theta_0$, and let an upgrade raise the bar to $\theta_1 > a_0$. Exposure stops and $a(s) = a_{\min} + (a_0 - a_{\min})e^{-\mu s}$. Restoring θ_0 after s rounds returns the skill to a_{prac} if and only if*

$$s \leq s^* = \mu^{-1} \log \frac{a_0 - a_{\min}}{\theta_0 - a_{\min}}; \tag{5}$$

later rollbacks restore the comparator but leave the skill in the low state.

Proposition 3 (Preservation and a uniqueness certificate)**.** *(a)* Weak interference. *For $a_{\min} < \theta_1 < a_{\max}$, an exact-estimate upgrade from a_{prac} to θ_1 causes entry only if $\varphi < (\theta_1 - a_{\min})/(a_{\max} - \theta_1)$; for fixed a_0 the window s^* grows as $1/\mu$, and with $\mu = 0$ an unexposed skill does not decay. (b)* Low-gain response. *For a continuously differentiable keep response $\rho(a)$, if $G = \sup_a \varphi\rho'(a)(a_{\max} - a) \leq 1$ the dynamics have a unique, globally attracting equilibrium.*

Part (a) protects an already competent learner, whereas part (b) certifies a unique equilibrium, which need not have high competence. The derivative certificate limits how strongly increased competence

amplifies exposure relative to decay. For a threshold CDF $F(a) = F_0((a-\theta)/s)$ with base density bounded by M, the bound $G \le \varphi(a_{\max} - a_{\min})M/s$ makes threshold concentration, rather than variance alone, the relevant quantity. Appendix E gives the exact fold envelope for the single log-link response: temperatures above Eq. (12) ensure uniqueness for every positive center. It also specifies when the envelope is attainable if centers are restricted. These results concern the scalar flow; full-relaxation iteration and neural parameter updates require separate analysis. Reducing interference also changes $\varphi = \kappa/\mu$, so it is not equivalent to reducing response gain at fixed learning parameters.

The propositions above assume constant interference μ. Appendix A.1 extends the persistence and recovery analysis to an interference rate $\mu(h)$ that decreases with accumulated exposure h, as suggested by our fits. During an interruption, h is fixed, so Eq. (5) holds at the current rate under the conditions of Proposition 5. Under that proposition's threshold condition, the two states persist, with the upper state approaching the fully practiced level. The exposure needed for recovery is bounded by the critical doses from Eq. (6) at the current and fully practiced rates.

Floors, margins, and noise. For $a_{\min} < \theta < a_{\text{prac}}$ and $a_0 < \theta$, a floor makes the low-state trajectory tend to $a^*(\delta)$ and cross θ in finite time exactly when

$$\delta > \delta_{\text{crit}} = \frac{\theta - a_{\min}}{\varphi(a_{\max} - \theta)}, \tag{6}$$

with crossing time $t_\delta = (\kappa\delta + \mu)^{-1} \log[(a^*(\delta) - a_0)/(a^*(\delta) - \theta)]$, which decreases in δ and diverges as $\delta \downarrow \delta_{\text{crit}}$. Training volume δt_δ and constant per-example spending also diverge at this boundary (Appendix F), so the minimum feasible dose need not be operationally attractive. These exact scalar statements concern threshold crossing from below; maintaining useful competence below the current comparator is a different objective. The dose depends on each region's (θ, φ), and an aggregate quota can be met by practiced regions while another receives nothing (Appendix G). The deployed router re-scores a fixed calibration set after every update, so $\hat{a}$ varies because the model changes, not because items are resampled. If this variation gives a deferred skill a keep probability of at least $\underline{p} > 0$ per round, it leaves the low state within $1/\underline{p}$ rounds in expectation; as a approaches $a_{\min}$ that probability can vanish and the low state becomes effectively absorbing. The same variation can also tip a kept skill below a nearby bar, so near the bar neither state is absorbing. Floors raise the escape probability by raising a, and a below-critical floor can succeed on a finite horizon (Appendix A).

Decoupling. Let per-round utility be $U_t = u_t a_t + (1 - u_t)(b - \lambda) - c(v_t)$ with nondecreasing training cost c, and suppose competence depends on training access only, $a_{t+1} = F(a_t, v_t)$.

Proposition 4 (Serving and learning separate)**.** *If u and v can be chosen independently, an optimal policy serves greedily, $u_t = \mathbf{1}\{a_t \ge b - \lambda\}$, and chooses $\{v_t\}$ to maximize $\sum_t [\max(a_t, b-\lambda) - c(v_t)]$. Handled-only learning ($v = u$) is weakly worse, and a floor that attains training dose δ below the bar pays $\delta(\theta - a_t)$ per round in serving utility that training access at the same dose avoids.*

Given $\{v_t\}$ the trajectory is fixed and the objective separates across rounds (Appendix A). The premise requires labels that do not depend on who served the query and comparable training costs $c(v)$ for the same dose. It separates serving from maintenance but does not solve the optimal training schedule. It fails when feedback arrives only on served answers or the student self-trains on its own outputs; those regimes are outside this result.

4 Experimental design

Overview. The main experiments follow the language-model student through a fallback upgrade, identify the training mechanisms, and compare ways to maintain competence. Appendix C contains their complete comparisons. The E1–E7 studies in Appendix I examine the model's response dynamics and complementary exposure settings; the bounded-memory image cascades in Appendix L test the principle in a different learning process.

Language-model cascade. The student is Qwen2.5-1.5B-Instruct with a LoRA adapter (Qwen Team, 2025; Hu et al., 2022). The adapter is updated on each round's training examples, and its weights carry forward to the next round. The fallback is Qwen2.5-7B-Instruct, replaced by Qwen2.5-14B-Instruct in the upgrade. Traffic comprises ten procedurally generated mathematics skills with 100

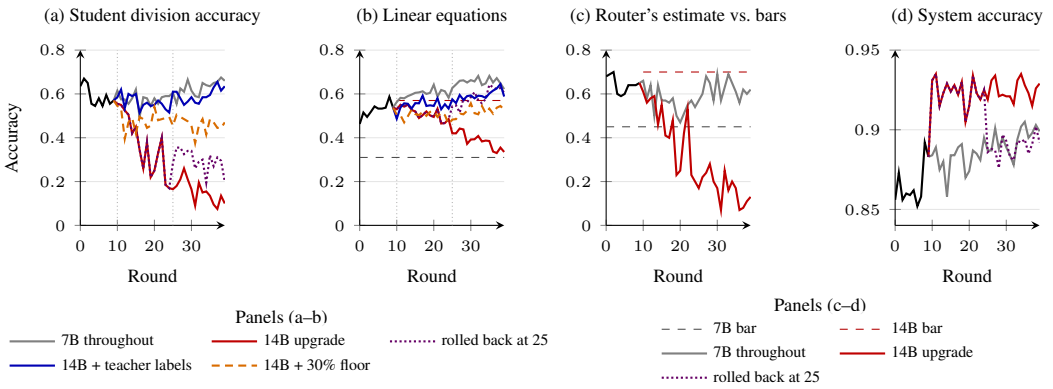

Figure 1: **A fallback upgrade reduces student accuracy while system accuracy rises.** Curves show matched intervention trajectories sharing a round-9 student checkpoint and incoming queries; Table 1 separately reports aggregate endpoints over eight seeds. (a) The upgrade at round 10 removes division training. Rollback at round 25 does not restore keeping in the plotted division trajectory; teacher labels preserve competence. (b) Linear-equation accuracy declines after the upgrade and improves after rollback. (c) The division control estimate remains between the original 0.45 bar and the upgraded 0.70 bar, while the upgraded estimate declines. (d) System accuracy rises despite the loss of student competence.

new queries per skill per round, including integer division, linear equations, and fraction reduction. The router applies Eq. (2) per skill, keeping or deferring all of a skill's queries, with $\hat{a}$ re-estimated on a fixed held-out calibration set after every update. This task-level router makes the link between assignment and exposure explicit; Section 6 also tests a smoother response. Student competence is measured on held-out evaluation queries excluded from training and router calibration; system accuracy, deferral fraction, and utility $U = A - \lambda C$ use the full evaluation split. During the upgrade the rate of malformed answers on division and linear equations stays at or below 0.5% in every arm, so the accuracy losses we report are wrong answers rather than format failures.

Paired interventions. Upgrade, recovery, row-restoration, and mechanism comparisons share their branch-point checkpoint and incoming queries. The assignment-by-access controls additionally hold every non-target batch fixed; matched-dose arms instead retain routing for other skills. Router variants repeat practice and upgrade from their own initial trajectories and are compared with their own pre-upgrade levels, rather than treated as matched checkpoint interventions. Protocols specify the intervention horizon and whether the endpoint is retained accuracy, routing, or serving utility.

Reporting. Late accuracy averages the last five evaluations, and $\pm$ values are standard errors over seeds (for a paired comparison, of the paired difference). Unless otherwise specified, experiments report means over eight seeds.

5 Result 1: Better system accuracy can mask student skill loss

Our experiments establish three results. First, a stronger fallback improves the cascade's answers while the student loses a deferred skill. Second, matched interventions identify target-task training access and interference from other updates as mechanisms of this loss. Third, continued training can preserve or recover competence while leaving the fallback available to serve queries. We develop these claims below; the appendix collects the full intervention sweeps and complementary tests.

The upgrade exposes a gap between serving performance and retained competence. System accuracy rises from 90.0% to 93.0%, while the student's division accuracy falls from 64.4% to 16.2% (Table 1). The fallback supplies better answers immediately, but the newly deferred skill loses its training stream. Linear equations show the same pattern; fraction reduction continues receiving examples and retains its accuracy.

Figure 1 follows this process over time. The higher routing bar removes division from student training while updates on other skills continue. The resulting decline matters for future deployment: a system

Deployment after round 10	Division	Linear eq.	Fraction red.	System acc.	Deferral
7B fallback throughout	0.644	0.655	0.881	0.900	0.100
14B upgrade	0.162	0.348	0.892	0.930	0.300
14B + teacher labels on deferred queries	0.652	0.615	0.870	0.916	0.187
14B + 30% exposure floor	0.475	0.531	0.900	0.907	0.203
14B, rolled back to 7B at round 25	0.489	0.625	0.893	0.895	0.133

Table 1: **Upgrade outcomes.** Means over eight seeds, each averaged over the final five rounds. These aggregate endpoints are distinct from the trajectories in Figure 1. Rollback at round 25 leaves division deferred on three of eight seeds.

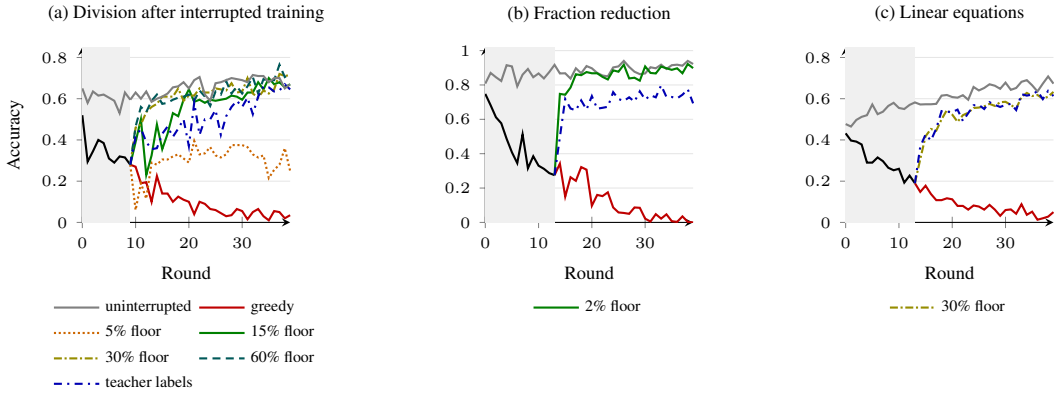

Figure 2: **Recovery after training interruptions.** Shading marks interruption; gray: uninterrupted twin; black: interruption; red: subsequent greedy routing; blue: teacher labels. (a) Division after ten interrupted rounds: the 5% floor retains accuracy 0.272 while remaining deferred; 15%, 30%, and 60% floors cross at update indices 6, 2, and 0. (b) Fraction reduction after fourteen interrupted rounds at $\lambda = 0$. (c) Linear equations interrupted for fourteen rounds alongside division, fraction reduction, and multiplication. Panels (b–c) show three-seed means. Table 5 gives endpoints; Table 6 gives division crossing counts over eight seeds.

can become more capable at answering its current traffic while its smaller model becomes less able to take that traffic back. Increased deferral also lowers routing utility at the specified fallback penalty, making system accuracy, student competence, and serving cost distinct deployment outcomes.

Preserving feedback changes this outcome. Teacher labels on deferred queries retain division accuracy at 65.2%, close to the unchanged-fallback control, and also preserve linear-equation performance. The stronger fallback can therefore contribute both answers and training feedback. Exposure floors and rollback recover substantial competence as well, with their full outcomes reported in Table 1. The next sections examine why these interventions work and when they restore student assignment.

The upgrade pattern extends to a second model family. A partial replication with OLMo-2 yields the same combination of improved system answers and reduced student division competence (Appendix C.6). Across three seeds, division accuracy under the upgraded fallback is below the unchanged-fallback control on every seed, with a paired mean loss of 0.133 ± 0.013, while system accuracy improves. Teacher labels and an exposure floor each recover part of the division loss. This experiment connects the central upgrade finding to a second family and provides further evidence that preserving feedback can help the student retain a newly deferred skill.

5.1 Persistence and recovery after an interruption

The loss persists because the routing decision removes the examples needed for recovery while other skills continue updating the same parameters. Interruption, rollback, and exposure interventions reveal how this feedback sustains a low state and how renewed practice restores the skill.

An interruption can outlast its cause. After training resumes under the same greedy rule, bar, and traffic as the uninterrupted twin, the interrupted division student remains far less accurate through the evaluation horizon (Figure 2). Fraction reduction and linear equations show similar

persistent gaps. The routing rule creates a feedback loop: a weakened student is deferred, and deferral withholds the target examples that could restore its competence. Lowering the bar or supplying training independently of assignment breaks this loop and brings division performance close to the uninterrupted twin (Table 5).

Recovery remains possible after rollback. Returning to the original fallback restores student assignment across many tested delays, including delays beyond the fitted scalar recovery windows. Some branches remain deferred, and repeated updates from the same stored state can produce different routing outcomes (Tables 3–4). The scalar model captures the exposure feedback and motivates intervention; these branch experiments show why an operator should also re-evaluate the updated student during recovery.

Prior practice helps the student withstand interruption. Division loses substantially less accuracy after in-loop practice, and students matched on competence and subsequent exposure follow similar short-run trajectories (Appendix C.2). Together, the interruption and rollback experiments motivate preserving training access before a skill falls below the routing bar and checking recovery against held-out competence.

6 Result 2: Training access and shared updates explain the loss

Target coverage repairs most of the upgrade gap. Replacing other-task rows with division examples at a fixed training budget raises division accuracy from 0.413 to 0.632, closing about 94% of the gap to the unchanged-fallback control (Table 7). Matching the control's non-target mixture or total row count provides much less recovery. This intervention directly tests the practical decision created by deferral: which examples continue to reach the student. Reallocating an existing row budget toward the deferred skill is enough to recover most of its lost accuracy in this comparison.

Serving and training can be separated experimentally. We next hold every non-target batch fixed and vary who answers division queries independently of whether the student trains on them. With true-labelled target examples, division accuracy rises from 0.485 to 0.611. Switching the serving model while holding training fixed produces bit-identical adapters (Table 8). Under this feedback protocol, the competence gain follows training access, and the fallback can continue answering throughout. This gives an empirical instance of Proposition 4 and a concrete maintenance option for a deferred skill.

Parameter isolation protects a skill during deferral. The frozen adapter retains its branch-point competence, whereas updates on other tasks reduce division accuracy. A per-skill adapter also preserves that competence between division updates (Figure 3). These controls identify interference from shared updates as a source of loss when target training stops. With target examples restored, both shared and isolated adapters learn; extended target training benefits the isolated adapter more in the tested comparison. The intervention therefore addresses both sides of maintenance: protecting a deferred skill from unrelated updates and enabling further improvement when its examples become available.

The exposure pattern also appears with a smooth task-level routing response. Uniform replay retains target examples and substantially limits the loss (Table 9). These variants connect the controlled mechanism tests to policies that maintain an ongoing supply of practice.

7 Result 3: Continued training preserves and restores competence

Small doses of practice can have substantial value. A 15% exposure floor restores division assignment on all eight test seeds; larger floors accelerate recovery (Table 6). In the separate fixed-quota sweep, the first 5% of exposure supplies about 59% of the full-exposure accuracy gain (Table 10). Together, these experiments show that useful maintenance can require only a fraction of the original training stream. The dose needed to regain assignment depends on the student's state and the routing bar, so retained accuracy and return to student serving should be assessed separately.

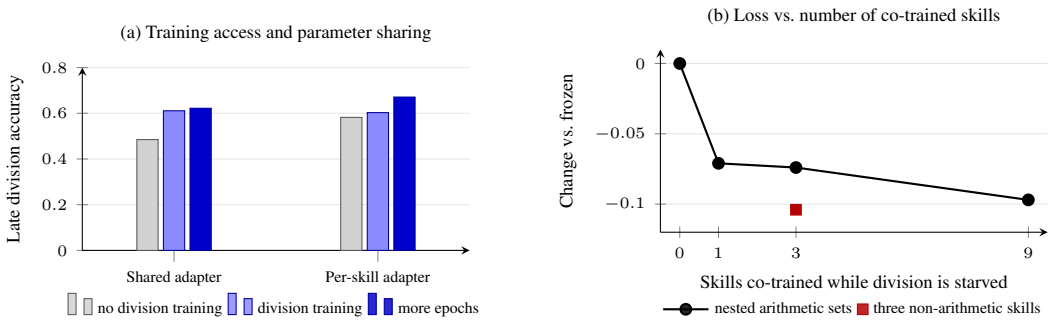

Figure 3: **Training access and parameter isolation preserve competence.** Matched interventions under the 14B fallback. (a) Accuracy with shared or per-skill adapters. (b) Loss relative to freezing as other skills train: addition/subtraction (K=1), plus multiplication and linear equations (K=3), all nine (K=9), or comparison, conversion, and primality (square). Paired differences ± SE: training access $+0.126 \pm 0.008$; freezing $+0.097 \pm 0.003$; per-skill versus shared with more epochs $+0.048 \pm 0.010$. Table 8 gives complete comparisons.

Recovery also improves serving outcomes. Exposure floors bring division competence close to the uninterrupted twin while reducing subsequent reliance on the fallback; teacher feedback reaches a similar utility to the twin (Table 5). Fraction reduction and linear equations also benefit from renewed exposure. The complete endpoints in the appendix distinguish the amount of competence restored from the cost and timing of returning queries to the student.

The exposure model guides the dose comparison. Using parameters fitted on separate identification seeds, Eq. (6) agrees with 34 of 40 seed–floor crossing outcomes. Every predicted crossing occurs, with additional recoveries at the smaller doses. This retrospective comparison supports the model's dose-dependent recovery relationship, while the branch-level variation motivates monitoring during deployment. Appendix C.2 reports the fitted parameters, crossing times, and noise-aware count intervals.

The fallback can answer while the student learns. The teacher-label result in Table 1 shows that competence can be maintained after serving shifts to the larger model. The assignment-by-access controls establish the corresponding result with true labels. Both require useful feedback to reach the student on deferred queries. Maintaining feedback as assignments change is also beneficial in the fraction-reduction experiment (Table 5).

Table 2 compares the resource requirements of four maintenance policies at a matched training dose. Independent training access and an exposure floor achieve similar accuracy, but independent access lets the fallback answer throughout. Replay attains the highest mean accuracy in this comparison using labels acquired previously. Teacher labels provide another source of feedback when fresh ground truth is unavailable. Thus the deployment choice can follow the available feedback: new labels, fallback-generated labels, or stored examples can each support continued learning.

These resource counts make the operational distinction explicit. An exposure floor couples student practice to student-served queries; independent access and replay let the operator choose serving and training separately. The comparison measures retained competence and query or label requirements. Selecting among these policies also requires accounting for feedback quality, update compute, and storage.

Complementary evidence from bounded-memory learners. Image cascades test the same exposure principle with students refit on finite memory windows. Exposure floors and replay improve worst-class retention on CRC100K and CIFAR-100, including with refreshed calibration. Under the harder separability-neutral shift, replay strengthens retention but the tested policies fall short of the joint retention-and-utility target. These experiments delimit how the principle transfers to a different learning process. Appendix L provides the datasets, shift conditions, calibration budgets, and complete comparisons.

Intervention	Student-answered queries	New labels	Late division
None (greedy; no dose)	0	0	0.415
Teacher labels on deferred queries	0	300 (teacher)	0.479
Exposure floor	294	294	0.495
Training access, fallback answers	0	300	0.502
Replay of stored examples	0	0 (previously acquired)	0.519

Table 2: **Query and label requirements for a 30% training dose** under the 14B fallback, ten rounds from a shared checkpoint. Means over eight seeds. Query and label counts cover division and are summed over the ten rounds. Other skills follow the router in this experiment; Figure 3 holds them with the student. The respective greedy baselines are 0.415 and 0.485. Paired differences from the floor $\pm$ SE: replay $+0.024 \pm 0.007$, training access $+0.007 \pm 0.02$, teacher labels -0.016 ± 0.03.

8 Implications and scope

Evaluate evolving decision-makers. A fixed-checkpoint comparison measures answers before routing induces different training histories. Training exposure and standalone per-skill competence should be tracked alongside serving accuracy and fallback use, and recovery should be checked against the operator's competence target rather than inferred from renewed assignment alone.

Match maintenance to the learning process. Where suitable feedback is available independently of serving, training access can preserve a deferred capability without changing serving assignments. Replay, teacher feedback, parameter isolation, and floors make different demands on labels, compute, memory, and adapter storage, so the reported routing utilities do not support a universal cheapest intervention.

Scope. Our language-model experiments use task-level routing and shared parameters, with training restricted to served queries in the baseline. Losses arise when updates on other tasks interfere with deferred skills; more prior practice, smaller updates, continued exposure, and parameter isolation limit this effect. Every served answer here receives its correct label. With noisier feedback or self-training on the student's own outputs, serving assignment could change the available labels and parameter updates; we do not test these regimes. Competence returns when useful exposure resumes, consistent with recoverable forgetting under fine-tuning (Kotha et al., 2024), while noisy decisions prevent exact branch forecasts from the scalar model. The detailed mechanism experiments use Qwen on procedural tasks, with the upgrade pattern partially replicated on OLMo-2 (Appendix C.6); the bounded-memory studies establish complementary exposure mechanisms under their feature, calibration, and shift assumptions. Generality to natural query streams and other model families, and the full costs of maintenance, remain to be evaluated.

9 Conclusion

Routing affects retained competence when it determines which examples reach the student. In our cascades, restoring target examples brings performance close to the control, and isolating task parameters prevents the tested interference. An exposure model gives conditions for unique equilibria and recovery. With feedback available independently of serving, training on deferred queries can preserve competence while the fallback answers. These findings support evaluating retained task performance alongside system accuracy and serving cost.

Reproducibility. Protocols, derivations, and complete endpoint tables are in the appendix. Code, generators, per-round histories, and the experimental settings will be released with the paper.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing writing, retrieval and discovery of related work, research ideation and execution, and proving mathematical claims. We have not used generative AI tools to draft sections of the paper or for generating synthetic datasets. We have reviewed all AI-assisted work. Specifically, LLM-assisted text polishing was carefully reviewed by the authors to ensure accurate representation of our ideas, literature retrieval and research ideation were validated through manual literature surveys, and all mathematical proofs generated or assisted by AI were rigorously verified step-by-step for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ali Aouad, Thodoris Lykouris, and Huiying Zhong. Human-AI productivity paradoxes: Modeling the interplay of skill, effort, and AI assistance. *arXiv preprint arXiv:2605.11350 (conference version: EC 2026)*, 2026.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Jonathan Bragg and Emma Brunskill. Fake it till you make it: Learning-compatible performance support. In *Uncertainty in Artificial Intelligence*, PMLR 115, 2020.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, 2025.

Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *arXiv preprint arXiv:1710.11214*, 2017. https://arxiv.org/abs/1710.11214.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

C K Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46, 1970.

Stephen Coate and Glenn C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, 1993.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905 (FAT/ML workshop)*, 2018.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. *arXiv preprint arXiv:2410.10347*, 2024.

Tong Ding et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 2025. https://arxiv.org/abs/2411.19666.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Dang Hoang Duy, Yannis Montreuil, Maxime Meyer, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Online learning-to-defer with varying experts. arXiv:2605.12340, 2026.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, 1999.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*, 2024.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Hoda Heidari, Michael Kearns, and Aaron Roth. Tight policy regret bounds for improving and decaying bandits. In *International Joint Conference on Artificial Intelligence*, 2016.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Lingxiao Huang and Nisheeth K. Vishnoi. Path dependence under adaptive AI delegation. *arXiv preprint arXiv:2603.02950v2*, May 7, 2026. https://arxiv.org/abs/2603.02950v2.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Wittawat Jitkrittum, Neha Gupta, Aditya K. Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, 2023.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo dataset, 2018. https://doi.org/10.5281/zenodo.1214456.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems*, 2017.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, 2020.

Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, and Marcello Restelli. Stochastic rising bandits. In *International Conference on Machine Learning*, 2022.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *International Conference on Machine Learning*, 2025.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Beyond augmented-action surrogates for multi-expert learning-to-defer. *arXiv preprint arXiv:2604.09414*, 2026a.

Yannis Montreuil, Letian Yu, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Learning to defer in non-stationary time series via switching state-space models. *arXiv preprint arXiv:2601.22538*, 2026b.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pp. 10520–10545, 2023.

Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online cascade learning for efficient inference over streams. In *International Conference on Machine Learning*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Vishakha Patil, Vineet Nair, Ganesh Ghalme, and Arindam Khan. Mitigating disparity while maximizing reward: Tight anytime guarantee for improving bandits. In *International Joint Conference on Artificial Intelligence*, pp. 4100–4108, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, 2020.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Qwen Team. Qwen2.5 technical report. arXiv:2412.15115, 2025.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, 2019.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, 2015.

Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Megha Srivastava, Erdem Biyik, Suvir Mirchandani, Noah Goodman, and Dorsa Sadigh. Assistive teaching of motor control tasks to humans. In *Advances in Neural Information Processing Systems*, 2022.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pp. 3475–3483, 2024.

Donald M Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, 2022.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11415–11434, 2023.

Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-ai dependence for learning to defer. In *International Conference on Machine Learning (ICML)*, 2024.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A): 287–298, 1988.

Yu Wu, Shuo Wu, Ye Tao, Yansong Li, and Anand D. Sarwate. From deferral to learning: Online in-context knowledge distillation for LLM cascades. *arXiv preprint arXiv:2509.22984*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations*, 2025.

A PROOFS FOR SECTION 3

The interval $[a_{\min}, a_{\max}]$ is forward invariant under Eq. (3): the vector field is nonnegative at its lower endpoint and nonpositive at its upper endpoint. For fixed exposure v, the solution from a_0 is

$$a(t) = a^*(v) + (a_0 - a^*(v))e^{-(\kappa v+\mu)t}, \tag{7}$$

which proves Eq. (4) and attraction to the constant-exposure equilibrium.

Proof of Proposition 1. Below θ, exposure is zero and $\dot{a} = -\mu(a - a_{\min})$. Since $a_{\min} < \theta$, trajectories starting below the threshold remain below it and converge to $a_{\min}$. At or above the threshold, exposure is one and $\dot{a} = (\kappa + \mu)(a_{\text{prac}} - a)$. Since $a_{\text{prac}} > \theta$, the trajectory stays at or above the threshold and converges to a_{prac}. Each equilibrium attracts its interval; the tie convention supplies the upper branch at $a = \theta$. □

Proof of Corollary 2. Because $\theta_1 > a_0$ and a does not increase without exposure, exposure stays zero during the upgrade and Eq. (7) with $v = 0$ gives $a(s) = a_{\min} + (a_0 - a_{\min})e^{-\mu s}$. When θ_0 is restored, the assumption $a_{\min} < \theta_0 < a_{\text{prac}}$ makes Proposition 1 applicable with threshold θ_0: the skill returns to a_{prac} if $a(s) \geq \theta_0$ and tends to $a_{\min}$ otherwise. Solving $a(s) \geq \theta_0$ gives Eq. (5). If $\theta_0 \geq a_{\text{prac}}$ there is no upper state to return to, and if $\theta_0 \leq a_{\min}$ every state returns. □

Proof of Proposition 3(a). From a_{prac}, an exact-estimate upgrade stops exposure only if $\theta_1 > a_{\text{prac}} = (a_{\min} + \varphi a_{\max})/(1 + \varphi)$, which rearranges to $\varphi < (\theta_1 - a_{\min})/(a_{\max} - \theta_1)$. For fixed a_0 and θ_0, the window s^* in Eq. (5) is proportional to $1/\mu$; if $a_0 = a_{\text{prac}}$, which itself rises as μ falls, it grows faster still. With $\mu = 0$ and $v = 0$ the right side of Eq. (3) vanishes, so every state is stationary without exposure. Part (b) is proved in Appendix E. □

Floor condition and crossing time. Below the routing threshold, a floor fixes exposure at $v = \delta$. A trajectory from $a_0 < \theta$ crosses in finite time exactly when $a^*(\delta) > \theta$, which rearranges to Eq. (6). If $a^*(\delta) = \theta$, Eq. (7) approaches θ without reaching it. When the inequality is strict, solving Eq. (7) for $a(t) = \theta$ gives the crossing time in Section 3.2; its monotonicity and logarithmic divergence are shown in Appendix F. If $\theta > a_{\text{prac}}$, even full exposure remains below the comparator: a floor then maintains competence without restoring keeping, as for the 30% floor and teacher labels under the 14B fallback in Figure 1, where the bar exceeds the practiced level.

Keep decisions under estimation noise. The deployed router re-scores the same calibration items after every update, so the estimate $\hat{a}_t$ is not a fresh binomial draw each round: it changes because the model changes. Let q_t denote the conditional probability, given the history, that a deferred skill is kept at round t. If $q_t \geq \underline{p} > 0$ for every round the skill remains deferred, the number of rounds until the first keep is stochastically dominated by a geometric variable with success probability $\underline{p}$, so the expected time to leave the low state is at most $1/\underline{p}$. The condition can fail: while the skill is deferred its competence falls toward $a_{\min}$, and if the calibration estimate falls far below the bar, q_t can become negligible, making the low state effectively absorbing on any practical horizon. This is the behavior of division, whose deferred estimates sit far below its bar; strongly fluctuating estimates can instead permit escape when they clear the bar. The same variation acts on a kept skill: if the bar is close to its competence, one update can move the estimate below the bar and stop exposure, which is the noise-triggered entry of Section 5. A kept skill remains kept after escape only if the full-exposure update that follows keeps its estimate above the bar; when competence then rises toward a_{prac} well above the bar, re-entry becomes unlikely. A floor raises deferred competence toward $a^*(\delta)$ and with it the escape probability, so on a finite horizon a below-critical floor can still succeed even though Eq. (6) fails.

Proof of Proposition 4. Fix any training-access sequence $\{v_t\}$. Because $a_{t+1} = F(a_t, v_t)$, the competence trajectory $\{a_t\}$ is determined by a_0 and $\{v_t\}$ and does not depend on $\{u_t\}$. The total utility is $\sum_t [u_t a_t + (1 - u_t)(b - \lambda)] - \sum_t c(v_t)$, and the first sum is maximized term by term by $u_t = \mathbf{1}\{a_t \geq b - \lambda\}$, giving $\max(a_t, b - \lambda)$. Maximizing over $\{v_t\}$ then gives the stated problem, which proves optimality of greedy serving. Handled-only learning adds the constraint $v_t = u_t$, so its optimum is no larger. A floor that attains dose δ while $a_t < \theta = b - \lambda$ sets $u_t = \delta$ and earns

$\delta a_t + (1-\delta)\theta$, whereas training access at the same dose sets $u_t = 0$ and earns θ; the difference is $\delta(\theta - a_t)$, and both policies generate the same competence trajectory. □

The proposition requires that the update dynamics depend on training access alone. Handled-only training imposes the constraint $u = v$; self-training on served outputs or a serving-dependent label process can instead violate the assignment-independent update premise; the assignment-by-access experiment of Section 6 tests the premise directly in our cascade.

A.1 Practice-dependent interference

The exposure-law fits favor, out of sample, a decay rate that falls with accumulated practice (Section 5.1). Let practice h accumulate exposure, $\dot{h} = v$, and replace μ in Eq. (3) by $\mu(h)$, continuous, positive, and nonincreasing, with $\mu_\infty = \lim_{h\to\infty} \mu(h) \geq 0$:

$$\dot{a} = \kappa\, v\, (a_{\max} - a) - \mu(h)\,(a - a_{\min}), \qquad \dot{h} = v. \tag{8}$$

Write $a^*(v;\mu) = (\mu a_{\min} + \kappa v a_{\max})/(\mu + \kappa v)$, $a_{\mathrm{prac}}(\mu) = a^*(1;\mu)$, and $\delta_{\mathrm{crit}}(\theta;\mu) = \mu(\theta - a_{\min})/[\kappa(a_{\max} - \theta)]$, which is Eq. (6) with $\varphi = \kappa/\mu$. Both $a^*(v;\mu)$ for $v > 0$ and $a_{\mathrm{prac}}(\mu)$ decrease in μ, while δ_{crit} increases in it. The fitted form $\mu(h) = \mu_0/(1 + h/h_0)$ has $\mu_\infty = 0$.

Proposition 5 (Practice-dependent interference)**.** *Under Eq.* (8) *with handled-only greedy routing and exact estimates, start from* (a_0, h_0). *(a)* Two states. *If* $a_{\min} < \theta < a_{\mathrm{prac}}(\mu(h_0))$*, states with* $a_0 < \theta$ *keep* $h = h_0$ *and converge to* $a_{\min}$. *States with* $a_0 \geq \theta$ *stay at or above* θ *and converge to* $a_{\mathrm{prac}}(\mu_\infty) \geq a_{\mathrm{prac}}(\mu(h_0))$. *(b)* Rollback window. *Corollary 2 holds exactly with* $\mu = \mu(h_0)$*, the rate at the upgrade, provided* $\theta_0 < a_{\mathrm{prac}}(\mu(h_0))$. *(c)* Floor bracket. *For* $a_0 < \theta < a_{\max}$ *and a floor* δ*, the skill crosses* θ *in finite time if* $\delta > \delta_{\mathrm{crit}}(\theta;\mu(h_0))$. *It then crosses no later than* t_δ *evaluated at* $\mu(h_0)$. *It never crosses if* $\delta \leq \delta_{\mathrm{crit}}(\theta;\mu_\infty)$.

Proof. The interval $[a_{\min}, a_{\max}]$ remains forward invariant, as for Eq. (3). (a) Below θ the exposure is zero, so h stays at h_0 and a obeys Eq. (3) with the constant rate $\mu(h_0)$; the proof of Proposition 1 applies. At or above θ the exposure is one, and $\dot{a} = (\kappa + \mu(h))(b(t) - a)$ with $b(t) = a_{\mathrm{prac}}(\mu(h(t)))$. Because h grows and μ is nonincreasing, $b(t)$ is nondecreasing, so $b(t) \geq b(0) > \theta$. Thus $\dot{a} > 0$ whenever $a = \theta$, and the trajectory cannot fall below θ. As $h = h_0 + t \to \infty$, $b(t) \to a_{\mathrm{prac}}(\mu_\infty)$. The relaxation rate $\kappa + \mu(h)$ is at least $\kappa > 0$, so $a(t) - b(t) \to 0$. (b) While the upgraded bar exceeds competence, the exposure is zero and h is frozen at its value at the upgrade. Eq. (8) then coincides with Eq. (3) at $\mu = \mu(h_0)$, which gives the decay path and Eq. (5). After θ_0 is restored, part (a) at the new state applies, because h has not changed. (c) With $v = \delta$ we have $\mu_\infty \leq \mu(h(t)) \leq \mu(h_0)$ for all t. For $a \geq a_{\min}$ this gives

$$\kappa\delta(a_{\max} - a) - \mu(h_0)(a - a_{\min}) \;\leq\; \dot{a} \;\leq\; \kappa\delta(a_{\max} - a) - \mu_\infty(a - a_{\min}).$$

By the scalar comparison theorem, $a(t)$ lies between the constant-rate solutions of Eq. (7) at $\mu(h_0)$ and at μ_∞ from the same a_0. If $\delta > \delta_{\mathrm{crit}}(\theta;\mu(h_0))$, the lower solution crosses θ at $t_\delta(\mu(h_0))$, and so the actual trajectory crosses no later. If $\delta \leq \delta_{\mathrm{crit}}(\theta;\mu_\infty)$, then $a^*(\delta;\mu_\infty) \leq \theta$. The upper solution then rises monotonically from $a_0 < \theta$ toward that limit without reaching θ, so the actual trajectory does not reach it either. □

The scalar model with constant μ is therefore the reference case, not a requirement. Practice lowers the critical dose and lengthens the rollback window, with the constant-rate values at the current practice level serving as conservative bounds. Two cases remain uncharacterized. When $a_{\mathrm{prac}}(\mu(h_0)) \leq \theta < a_{\mathrm{prac}}(\mu_\infty)$, a kept skill survives only if practice accumulates quickly enough. And the floor's exact critical dose inside the bracket depends on the whole path of h. With $\mu_\infty = 0$ the lower end of the bracket is zero: in this model, any positive floor eventually drives $a^*(\delta;\mu(h))$ above the bar, and so rescues the skill if it runs long enough.

B Language-model protocol

The student uses LoRA rank 16, scale 32, and dropout 0.05 on all attention and MLP projections. Bootstrap learning uses learning rate 10^{-4}; round updates use AdamW at 2×10^{-4} for two epochs,

batch size 8, bf16 precision, and maximum sequence length 512. Optimizer state restarts each round while adapter weights carry forward; extended-training arms use six epochs. The fallback penalty is $\lambda = 0.25$, including both models in the upgrade, 0 for the fraction-reduction interruption, 0.20 for novice acquisition, and 0.60 for the higher-penalty recovery intervention. The multi-target study prices fraction reduction at 0.10 before its zero-penalty follow-up.

Data and tasks. For each task, procedural generators written for this study produce 100 bootstrap examples, 100 calibration examples, 200 evaluation examples, and 100 feedback examples for each of 40 rounds; streams are separated by rejection sampling. The generators are modeled on the DeepMind Mathematics module taxonomy but are not its released generator. The ten tasks are addition/subtraction, multiplication, integer division, single-variable linear equations, pairwise comparison, unit conversion, greatest common divisor, primality, addition of linear polynomials, and fraction reduction. Integer division samples divisor and quotient from $\{2, \ldots, 500\}$ and asks for the quotient of their product and divisor. Linear equations sample $a \in \{1, \ldots, 9\}$, $b \in \{1, \ldots, 30\}$, and integer solution $x \in \{1, \ldots, 200\}$. Fraction reduction asks for the probability of drawing a red ball from a bag, as a reduced fraction with denominator at most 100,000.

Training and inference. The practiced student is bootstrapped on labelled examples of every task; a novice student receives only a brief format warm-up. After each round the LoRA adapter is updated on that round's training rows; adapter weights carry over between rounds. Only answer tokens contribute to the loss. Inference uses greedy decoding; fallback outputs are cached for the query stream. Teacher-label arms train on the fallback's generated answers, errors included.

Routing and measurement. The router recomputes student accuracy on the same 100 calibration examples after every update; fallback calibration accuracy is fixed for each teacher on the calibration set. Penalties are fixed per task within each experiment. Floors keep each otherwise-deferred example independently with probability δ, with common random numbers so that lower-floor kept sets nest within higher-floor sets. Evaluation uses normalized exact match on 200 held-out examples per task. A task is *kept* if its estimate clears the bar after each of the last three updates and *deferred* if it falls below after all three.

Experiment families. *Upgrade.* Ten practiced rounds under the 7B fallback, then five branches from the shared round-9 checkpoint: 7B throughout; 14B from round 10; 14B with a 30% floor; 14B with ground truth on kept queries and 14B answers on deferred ones; 14B until round 24 and 7B from round 25. Fraction reduction is never deferred in this experiment.

Interruption. From the bootstrapped checkpoint, the target task is routed to the fallback for ten (division) or fourteen (fraction reduction) rounds while other tasks keep training; greedy, floor, and teacher-label arms then run to round 39. The uninterrupted twin is practiced in the loop from the same checkpoint.

Lower the bar. From the interrupted checkpoints, the fallback is replaced by Qwen2.5-0.5B or the penalty is raised; both lower every task's bar below the starved student and coincide.

Rollback timing. Rollbacks to the 7B fallback are branched from stored upgraded checkpoints at rounds 11–24 and run ten further rounds.

Row restoration. From the upgrade branch point, the student is retrained on the upgraded branch's exact per-round rows, with division's rows returned in place of other rows (same count), with the control's non-target mixture, or padded with other rows to the control's count; a transfer test adds one 100-row block (division, linear equations, or primality) for two updates from a starved checkpoint.

Assignment by access. From the round-9 checkpoint under the 14B fallback, division's assignment and training access are set independently for ten rounds while every non-target task is kept with the student, so non-target batches are byte-identical across cells. The same design supplies the frozen adapter, the co-training sets, and the per-skill adapter (forked at the branch point and trained only on division rows, with standard or extended training), while the matched-dose interventions at 30% of division's traffic allow non-target tasks to follow the router.

Response variants. Practice and upgrade are re-run with 10% uniform replay, a probabilistic task-level router, or a query-level router that keeps individual queries according to the student's answer log-probability (whose AUROC for division correctness is 0.80–0.89).

Reporting. Kept and deferred states refer to the specified evaluation horizon. Figure 1 displays matched intervention trajectories, whereas Table 1 reports eight-seed aggregate endpoints. For teacher-labelled division, the plotted final-five mean is 0.626 and the table aggregate is 0.652. Terminal and time-averaged endpoints can differ: in Figure 1, time-averaged division accuracy over rounds 10–39 is 0.262 against a late-five accuracy of 0.108, and it is still declining at the horizon. The malformed-answer rate, the share of outputs the scorer cannot parse, stays at or below 0.5% for the target skills in every upgrade arm.

C Complete language-model results

C.1 Upgrade and rollback

Rollback timing. Table 3 reports timed rollback interventions after an upgrade raises the division bar from 0.40 to 0.60, and a second sequence in which it rises from 0.45 to 0.70. Recovery is assessed from the calibration estimate used by the router, with held-out accuracy reported separately. A 15% floor applied from the round-18 and round-24 states crossed the bar after two updates and one update and ended at 0.595 and 0.56. For seed 1400, the round-18 rollback restores keeping after the calibration estimate rises from 0.34 to 0.42 on the first update, with held-out accuracy reaching 0.57 ten rounds later. The original round-20, round-21, round-23, and round-24 branches remain deferred; one of three optimizer-seed reruns of the round-21 state restores keeping (Table 4). These branches illustrate opportunities for recovery after rollback as the updated student is re-evaluated. On the second sequence all fourteen branches, including ones starting at 0.23, crossed the bar within four updates. With the identification-seed fit of Section 7, Eq. (5) gives $s^* = 2.5$ (90% seed-bootstrap interval 1.5–3.7) and 2.8 (1.7–4.2) upgraded rounds for the two sequences. It uses each sequence's router estimate at the upgrade, 0.55 and 0.65, and its 7B bar. Using the fitted decay rate after ten rounds of practice (0.087 per round) gives rollback windows of about 4.5 and 5 rounds. Both sequences recovered for longer, through eight and at least fourteen upgraded rounds.

Rollback timing		Router estimate		Division accuracy		
Round	Delay	Before	After 1 update	Before	After 10 rounds	Routing
Seed 1400	Routing bar: $0.60 \to 0.40$					
11–16	1–6	0.460–0.530	0.480–0.550	0.530–0.630	0.560–0.640	Kept
18	8	0.340	0.420	0.405	0.570	Kept
20	10	0.350	0.390	0.350	0.310	Deferred
21	11	0.370	0.350	0.450	0.195	Deferred*
23	13	0.350	0.290	0.460	0.275	Deferred
24	14	0.290	0.290	0.355	0.190	Deferred
Seed 1700	Routing bar: $0.70 \to 0.45$					
11–24	1–14	0.230–0.610	0.430–0.670	—	—	Kept†

Table 3: **Recovery after fallback rollback.** Delay counts upgraded rounds. Router estimates use calibration queries; accuracy uses held-out evaluation queries. Ranges summarize grouped branches. *One of three round-21 reruns returns to keeping (Table 4). †Thirteen branches return after one update and one after four.

Escape as a rate. The escape-rate test reran eight stored starved upgrade states with the 7B bar restored. Each state used three optimizer seeds, with its feedback stream, calibration and evaluation sets, and training rows held fixed (Table 4). The states, predictions, and pass criteria were fixed in a ledger before the runs. First-update escape probability increases with the signed margin $\hat{a} - \theta$: a logistic fit has slope 11.6 (90% bootstrap 3.0–29.6) and midpoint -0.064. The state-level outcomes vary around that trend: all three runs escape at margins of -0.05 and -0.11, while none escape at -0.08 or -0.14. Across the eight states, 8 of 24 runs escape on the first update and 9 of 24 escape

within the evaluated runs, whose horizons are specified in Table 4. From the state 0.29 below its bar, none of three runs escaped without a floor, while 5% and 15% floors escaped on all three seeds.

Stored state (seed, round)	Margin $\hat{a} - \theta$	Escaped at first update	Escaped within the run
1400, r21	-0.03	1/3	1/3
1700, r18	-0.05	3/3	3/3
1200, r25	-0.08	0/3	0/3
1600, r26	-0.11	3/3	3/3
900, r31	-0.14	0/3	0/3
1300, r33	-0.18	0/3	0/3
1700, r27	-0.23	1/3	2/3 (five rounds)
1700, r34	-0.29	0/3	0/3

Table 4: Escape-rate replicates: each stored starved state rolled back to the 7B bar under three optimizer seeds with identical data. Runs are two rounds except the two deepest states, which run five. The 1400/r21 state is the round-21 branch of Table 3. Fitted first-update escape probability increases with the signed margin $\hat{a} - \theta$ (logistic slope 11.6, midpoint -0.064).

C.2 Interruption suites

Task	Intervention	Late student accuracy	Utility	Deferral
Division	Uninterrupted twin	0.680 ± 0.008	0.881 ± 0.006	0.100 ± 0.012
Division	Greedy	0.034 ± 0.006	0.845 ± 0.012	0.200 ± 0.025
Division	Floor 5%	0.272 ± 0.031	0.832 ± 0.015	0.187 ± 0.022
Division	Floor 15%	0.675 ± 0.014	0.868 ± 0.009	0.077 ± 0.008
Division	Floor 30%	0.683 ± 0.011	0.873 ± 0.007	0.061 ± 0.007
Division	Floor 60%	0.713 ± 0.009	0.884 ± 0.005	0.035 ± 0.004
Division	Teacher labels	0.649 ± 0.022	0.880 ± 0.008	0.100 ± 0.015
Division	Lower the bar	0.708 ± 0.015	—	—
Fraction reduction	Uninterrupted twin	0.921 ± 0.008	0.883 ± 0.007	0.067 ± 0.009
Fraction reduction	Greedy	0.010 ± 0.006	0.852 ± 0.014	0.167 ± 0.018
Fraction reduction	Floor 2%	0.895 ± 0.025	0.886 ± 0.005	0.032 ± 0.005
Fraction reduction	Floor 5%	0.904 ± 0.027	0.887 ± 0.006	0.062 ± 0.008
Fraction reduction	Floor 10%	0.916 ± 0.016	0.890 ± 0.004	0.057 ± 0.007
Fraction reduction	Floor 30%	0.897 ± 0.022	0.884 ± 0.007	0.044 ± 0.006
Fraction reduction	Teacher labels	0.733 ± 0.157	0.873 ± 0.011	0.100 ± 0.014
Fraction reduction	Fallback's answers only	0.422 ± 0.050	0.801 ± 0.021	0.133 ± 0.022
Linear equations*	Uninterrupted twin	0.671 ± 0.010	—	—
Linear equations*	Greedy	0.033 ± 0.005	—	—
Linear equations*	Floor 30%	0.615 ± 0.036	—	—
Linear equations*	Teacher labels	0.621 ± 0.010	—	—

Table 5: Interruption interventions corresponding to Figure 2. Accuracy is the late-five mean; utility and deferral are terminal. Table 6 gives crossing outcomes for all eight test seeds. Fraction-reduction and linear-equation accuracies are mean ± SE. *Interrupted for fourteen rounds together with division, fraction reduction, and multiplication; system-level utility and deferral for that experiment are given in the text below.

Testing the critical dose. The scalar law's parameters are fitted on the four identification seeds only, never on the test seeds scored here. The equilibrium $a^*(v)$ of Eq. (4) is fitted to the late means of the fixed-quota dose sweep. The decay rate is fitted to the four starved-division trajectories of the confirmation run, with $a_{\min}$ fixed from the equilibrium fit. This gives $a_{\min} = 0.10$, $a_{\max} = 0.56$, $\varphi = 21.8$, $\mu = 0.16$ and $\kappa = \varphi\mu = 3.5$ per round, with 90% intervals from 2,000 resamples of whole seeds: $a_{\min}$ 0.09–0.11, $a_{\max}$ 0.55–0.57, φ 18.3–26.1, μ 0.11–0.27. The bootstrap intervals summarize variation across the four identification seeds. For each of the eight test seeds of the division suite we take its bar θ (0.33–0.49) and its router estimate at the start of the floor as a_0, then compute δ_{crit} and t_δ from Section 3.2. A predicted crossing time t_δ corresponds to the zero-based update index $\lceil t_\delta \rceil - 1$, the convention of Figure 2a. The critical doses range from 0.05 to 0.25 across seeds. All predicted crossings occur, with six additional observed recoveries at the smaller doses (Table 6). This check was run after the suite's outcomes were known; the parameters, however, come only from identification data.

	Observed		Eq. (6) and t_δ		
Floor	Crossed	Median update index	Crossed (90%)	Median update index	Agreement
0%	0/8	—	0/8 (0–0)	—	8/8
5%	5/8	2	1/8 (0–1)	8	4/8
15%	8/8	2	6/8 (5–7)	2	6/8
30%	8/8	0.5	8/8 (7–8)	0.5	8/8
60%	8/8	0	8/8 (8–8)	0	8/8

Table 6: Division floors after interruption across the eight test seeds, against the scalar law fitted on identification seeds. "Crossed" counts seeds whose router returns division to keeping within thirty rounds. The predicted count's interval is over seed-bootstrap parameter draws. Where both cross, the mean absolute error in the crossing index is 1.8, 1.1, and 0 at 15, 30, and 60%. The preregistered noise-aware ledger for this suite gave 90% count intervals of 0–1, 4–8, 6–8, 6–8, and 6–8.

Effect of prior training. When the same interruption follows ten rounds of in-loop practice rather than the bootstrap, division falls only from 0.61 to 0.53 over the interruption instead of from 0.52 to 0.28, and the student's estimate never drops below the routing bar. A scalar competence law whose decay rate falls with accumulated exposure, $\mu(h) = \mu_0/(1 + h/h_0)$ (0.254 per round at a fresh checkpoint, 0.087 after ten in-loop rounds), captures the measured exposure dependence (held-out RMSE 0.115, 90% interval coverage 0.855). Students matched on competence but reached through different histories and then given byte-identical exposure diverge by at most 0.035 over eight rounds. The practiced checkpoint retains more competence during interruption, and matching competence and subsequent exposure yields similar trajectories across histories.

C.3 Decomposition and interference

Training rows, rounds 10–24, under the 14B fallback	Division, rounds 20–24
7B control	0.645
Upgraded rows, division's rows returned (same total)	0.632
Upgraded rows (division starved)	0.413
Upgraded rows with the control's non-target mixture	0.367
Upgraded rows padded to the control's count	0.330
Two updates from a starved checkpoint, gain over adding nothing	
100 division rows	$+0.27$
100 linear-equation rows	$+0.07$
100 primality rows	-0.07

Table 7: Returning training rows at a fixed budget (upper block) and illustrative transfer effects of adding a 100-row block for two updates from a starved checkpoint (lower block).

C.4 Update size

The Qwen experiments use the protocol in Appendix B: Extended-training arms increase the number of epochs.

C.5 Response variants and several starved skills

When division, fraction reduction, linear equations, and multiplication are interrupted together for fourteen rounds, post-interruption greedy routing leaves division and linear equations deferred. For linear equations, late-five accuracy is 0.033 ± 0.005 under greedy routing, 0.615 ± 0.036 with a 30% floor, and 0.621 ± 0.010 with teacher labels, compared with 0.671 ± 0.010 for the uninterrupted twin (Table 5). The 30% floor returns both division and linear equations to keeping; teacher labels also restore their competence. Fraction reduction escaped the low state on its own after the interruption. Multiplication's bar sits just under its practiced accuracy, so it is often deferred even without interruption; it is not counted as a target.

Arm (ten rounds under the 14B fallback)	Late division
Fallback answers, no division training	0.485
Student answers, no division training	0.485 (bit-identical adapter)
Fallback answers, division training (true labels)	0.611
Student answers, division training (true labels)	0.611 (bit-identical adapter)
Fallback answers, division training (teacher labels)	0.576
Frozen adapter (= per-skill adapter, no training)	0.582 (no change from the branch point)
Optimistic (UCB) exploration, student answers 67%	0.585
Shared adapter, more epochs	0.622
Per-skill adapter, division training	0.603
Per-skill adapter, division training, more epochs	0.671
Per-skill adapter, replay only, more epochs	0.598 (peaks at 0.665, ends at 0.580)

Table 8: Assignment, training access, and interference, means over seeds. Every non-target batch is identical across arms, and other skills are held with the student. Relative to the frozen adapter, co-training addition/subtraction costs 0.071; adding multiplication and linear equations 0.074; comparison, conversion and primality 0.104; all nine 0.097. Paired differences ± SE: teacher minus true labels -0.034 ± 0.015; UCB minus no training $+0.100 \pm 0.020$; more epochs $+0.012 \pm 0.024$ for the shared and $+0.068 \pm 0.014$ for the per-skill adapter; per-skill replay-only with more epochs minus shared with true labels -0.013 ± 0.023; co-training contrasts -0.071 ± 0.011 (K 0 to 1), -0.004 ± 0.009 (1 to 3), -0.022 ± 0.009 (3 to 9), and arithmetic versus non-arithmetic at K=3 $+0.030 \pm 0.044$. The per-skill adapter with more epochs uses about a third of the shared adapter's gradient steps.

		Rounds 20–24		
Variant	Before upgrade	14B upgrade	7B control	Division training after the upgrade
Sharp task-level router	0.607	0.413	0.645	none
Smooth task-level router	0.581	0.349	0.651	none
10% uniform replay	0.581	0.542	—	about 6 replayed rows per round

Table 9: Response variants repeat the practice and upgrade schedule. The routers have matched 7B controls from the same round-9 checkpoint; replay is compared with its own pre-upgrade level (rounds 5–9).

C.6 The upgrade on a second model family

We repeated the upgrade of Table 1 with OLMo-2 models: an OLMo-2-0425-1B-Instruct student with a LoRA adapter, an OLMo-2-1124-7B-Instruct fallback, and an OLMo-2-1124-13B-Instruct upgrade, on the same seeds (900, 1500, 1700). Every OLMo run departs from the Qwen protocol in two ways. It trains and routes three skills: division, linear equations, and fraction reduction. It also uses $\lambda = 0.30$, set from a calibration probe on seed 1700, one of the reported seeds; at 0.25 the practiced division estimate is already below the 7B bar on two seeds. The pre-registered criterion required division under the upgrade to end at least 0.10 below the 7B control, with the paired difference positive on every seed, system accuracy to rise, and teacher labels and a 30% floor each to keep division above the upgraded arm.

The reported run uses an upgrade after three practiced rounds, a horizon of fifteen rounds, and a round learning rate of 1×10^{-4}, with no rollback arm. Division is deferred from the first or second post-upgrade update on every seed and falls from 0.201 to 0.068 (paired loss 0.133 ± 0.013; 0.157, 0.111 and 0.132 by seed), while system accuracy rises by 0.12. Teacher labels and the floor recover part of the loss, providing a further example of preserving training access alongside a fallback upgrade.

D Routing calibration and the feedback map

This appendix develops the analytical results in the notation of the main text. Throughout, exposure means access to training for the adaptive decision-maker. In the language-model and image cascades this is the student, so ρ is a *keep* probability. In a conventional L2D system with an adaptive external expert it would instead be that expert's deferral probability. The scalar dynamics depend on exposure, while the cost sign depends on which branch incurs the charge.

Division exposure	0%	5%	10%	15%	20%	30%	45%	60%	80%	100%
Late division accuracy	.095	.384	.419	.435	.482	.485	.505	.545	.587	.581

Table 10: Division dose response: fixed per-round division quotas, other skills trained as usual. Accuracy rises by 28.9 percentage points between 0% and 5% exposure and by a further 19.7 points between 5% and 100%.

D.1 Immediate utility and calibrated scores

For a fixed input or observable stratum, the student's expected utility is a and the fallback's is $b - \lambda$. Comparing the two proves Eq. (2). When $b - \lambda \leq 0$, the nonnegative-accuracy student is always weakly preferred. When $a > 0$ and $\theta = b - \lambda > 0$, a two-action weighted cross-entropy

$$L(s; a, \theta) = -a \log s_{\text{keep}} - \theta \log s_{\text{defer}}, \qquad s_{\text{keep}} + s_{\text{defer}} = 1,$$

has minimizer $s^* = (a, \theta)/(a + \theta)$. A Lagrange multiplier, or the nonnegativity of KL divergence, gives the minimizer. Thus its log-score difference is $\log a - \log \theta$, and choosing its larger component recovers the utility comparison. This is a population calibration statement, distinct from how the main experiments estimate regional accuracies.

For the standard $K + 1$ classification formulation, let the fixed predictor have posteriors $\eta_y(x)$ and the adaptive expert have accuracy $a(x)$ at cost $c \geq 0$. The utility comparison is $a - c \geq \eta_{\max}$, with threshold $\theta = \eta_{\max} + c$. The weighted loss

$$L(s; x) = -\sum_{y=1}^{K} \eta_y(x) \log s_y - (a(x) - c)_+ \log s_\perp$$

has closed-simplex minimizer proportional to $(\eta_1, \ldots, \eta_K, (a - c)_+)$. For positive weights, finite softmax logits attain it; zero weights are attained as limits in the open softmax parameterization. The largest-score rule therefore gives the Bayes comparison. For $c < a_{\min}$ the relevant statistic is $\log(a - c) - \log \eta_{\max}$. This is the surrogate studied by the supplementary learned-head experiment. Here c charges the adaptive branch; the main cascade's λ charges the fallback branch. They are separate conventions with the same exposure dynamics once θ is specified.

Logistic noise on a calibrated log-score difference yields the response

$$\rho_T(a) = \sigma\left(\frac{\ell(a) - \ell(\theta)}{T}\right), \qquad \sigma(z) = (1 + e^{-z})^{-1}, \quad T > 0. \tag{9}$$

The noise model is an additional assumption. Calibration identifies the statistic; it does not require logistic noise. The log link $\ell(a) = \log a$ corresponds to the two-action comparison; $\ell_c(a) = \log(a - c)$ corresponds to the positive cost-adjusted expert weight. We also analyze the probability-logit link $\log(a/(1 - a))$ as an alternative response family.

D.2 Equilibrium sensitivity and retraining

For constant exposure, Eq. (4) has derivative

$$\frac{da^*(u)}{du} = \frac{\varphi(a_{\max} - a_{\min})}{(1 + \varphi u)^2}, \qquad S := \sup_{u \in [0,1]} \frac{da^*(u)}{du} = \varphi(a_{\max} - a_{\min}). \tag{10}$$

The largest sensitivity occurs at zero exposure. Under *full-relaxation retraining*, the router supplies a response, competence equilibrates under that response, and the resulting competence is used for the next routing update. This gives the map $\Phi(a) = a^*(\rho(a))$. If ρ is L_ρ-Lipschitz and $SL_\rho < 1$, then Φ is a contraction of the competence interval. It has one fixed point and the iterates converge geometrically. This follows directly by composing the two Lipschitz bounds and applying the contraction theorem.

A learned router adds sensitivity through its optimization. Suppose its population risk $R(v; a)$ is γ-strongly convex in router parameters v, has cross-derivative norm bounded by β, and has a unique differentiable minimizer $\hat{v}(a)$. Differentiating $\nabla_v R(\hat{v}(a); a) = 0$ gives

$$\hat{v}'(a) = -[\nabla^2_{vv} R]^{-1} \nabla^2_{va} R, \qquad \|\hat{v}'(a)\| \leq \beta/\gamma.$$

If the deployed response is L_v-Lipschitz in v, the sufficient contraction condition is $SL_v\beta/\gamma < 1$. This is an alternative way to bound L_ρ, not an additional factor to multiply into a directly measured response derivative. Failing either sufficient bound leaves uniqueness unresolved.

Continuous time and finite rounds. The flow and full-relaxation iteration share fixed points, but their convergence conditions must be checked separately. For sharp routing, an Euler step with $0 < h \leq 1/(\kappa + \mu)$ is a convex combination of the current state and the relevant constant-exposure equilibrium on each branch. Under Proposition 1, each side of θ is therefore invariant and the discrete iteration preserves the two basins. Larger steps can cross a separatrix. The neural-network experiments use actual parameter updates, not Euler steps or full relaxation.

E SMOOTH ROUTING AND HETEROGENEOUS THRESHOLDS

Write the drift in units of μ as

$$H(a) = \varphi\rho(a)(a_{\max} - a) - (a - a_{\min}), \qquad \dot{a} = \mu H(a).$$

Boundary signs ensure an equilibrium. The following results characterize how the shape of ρ changes its uniqueness and multiplicity.

E.1 PROOF OF PROPOSITION 3(B)

Differentiating gives

$$H'(a) = \varphi\rho'(a)(a_{\max} - a) - \varphi\rho(a) - 1.$$

Under $G \leq 1$, this derivative is nonpositive, so H is nonincreasing. If it had two distinct zeros it would vanish throughout the interval between them. On any interior part of such an interval,

$$\rho(a) = \frac{a - a_{\min}}{\varphi(a_{\max} - a)}, \qquad \varphi\rho'(a)(a_{\max} - a) = \frac{a_{\max} - a_{\min}}{a_{\max} - a} > 1,$$

contradicting the premise. The root is unique, and the one-dimensional sign pattern makes it globally attracting. The proof extends to absolutely continuous responses when the derivative bound holds almost everywhere.

E.2 THRESHOLD DISPERSION

Let case-level thresholds have CDF F, giving $\rho(a) = F(a)$. If $F(a) = F_0((a - \theta)/s)$ and its base density satisfies $\operatorname{ess\,sup} f_0 = M < \infty$, then

$$G \leq \frac{\varphi(a_{\max} - a_{\min})M}{s}.$$

Thus $s \geq \varphi(a_{\max} - a_{\min})M$ certifies uniqueness for every center θ. For Gaussian thresholds, $M = 1/\sqrt{2\pi}$; for logistic thresholds, $M = 1/4$. The certificate controls density concentration. A variance statistic alone does not bound that concentration: a distribution can retain a narrow central spike while placing enough mass in its tails to maintain a fixed positive variance.

For Gaussian thresholds at a specified center, the exact derivative-bound check maximizes

$$g(a) = \frac{\varphi(a_{\max} - a)}{s\sqrt{2\pi}} \exp\left(-\frac{(a - \theta)^2}{2s^2}\right).$$

Its interior maximizer is

$$a^\dagger = \frac{\theta + a_{\max} - \sqrt{(a_{\max} - \theta)^2 + 4s^2}}{2},$$

clipped to $[a_{\min}, a_{\max}]$. This follows from $g'/g = -(a - \theta)/s^2 - 1/(a_{\max} - a)$. Evaluating g only at the density peak $a = \theta$ can underestimate the feedback gain, because the remaining learning gap is larger below that point. A failed derivative certificate does not establish multiple equilibria; direct root analysis can still show uniqueness.

When $a_{\min} < \theta < a_{\text{prac}}$, Gaussian responses with sufficiently small s have multiple equilibria. Choose points below and above the threshold where the sharp drift is respectively negative and positive, and combine them with the boundary signs. Away from θ, the smooth drift converges to the sharp drift as $s \to 0$, preserving these signs. Continuity gives three intervening roots. Distinct competence states therefore persist under sufficiently concentrated smooth thresholds.

E.3 Exact fold locus and log-link envelope

For Eq. (9), let $r(a) = (a - a_{\min})/[\varphi(a_{\max} - a)]$. At an equilibrium, $\rho_T(a) = r(a)$; an interior tangency also satisfies $H'(a) = 0$. Since $\rho_T' = \rho_T(1 - \rho_T)\ell'/T$, eliminating the response gives

$$T(a) = \ell'(a)\frac{(a - a_{\min})[\varphi(a_{\max} - a) - (a - a_{\min})]}{\varphi(a_{\max} - a_{\min})}, \quad a \in (a_{\min}, a_{\text{prac}}). \tag{11}$$

The center is recovered by

$$\theta(a) = \ell^{-1}\left(\ell(a) - T(a)\log\frac{r(a)}{1 - r(a)}\right).$$

These equations specify every tangency in the link family. For $\ell = \log$ with $0 < a_{\min} < a_{\max}$, maximizing Eq. (11) yields

$$T_c(\varphi) = \frac{\varphi(a_{\max} - a_{\min})}{\left[\sqrt{\varphi a_{\max} + a_{\min}} + \sqrt{(\varphi + 1)a_{\min}}\right]^2}, \quad a_{\text{tip}} = \sqrt{a_{\min}a_{\text{prac}}}. \tag{12}$$

To verify the maximization, write the locus as a positive constant times $(a_{\min} + a_{\text{prac}}) - a - a_{\min}a_{\text{prac}}/a$. Its derivative vanishes at the displayed geometric mean and its second derivative is strictly negative.

For $T > T_c$, there is a unique equilibrium for every positive center. One direct proof compares $\ell(a) - T\log(r(a)/(1 - r(a)))$ to the constant $\ell(\theta)$ on $(a_{\min}, a_{\text{prac}})$. Its derivative has the sign of $T(a) - T$, so it is strictly decreasing above the envelope and runs from $+\infty$ to $-\infty$. The envelope increases with φ and has limit

$$T_{\text{sat}} = \frac{\sqrt{a_{\max}} - \sqrt{a_{\min}}}{\sqrt{a_{\max}} + \sqrt{a_{\min}}}.$$

For the shifted-log link $\log(a - c)$, substitute $(a_{\min} - c, a_{\max} - c)$ into Eq. (12) and translate the maximizing state back by c. The assumption $c < a_{\min}$ keeps the link defined. The probability-logit link uses $\ell' = 1/[a(1 - a)]$ in the general locus, with $0 < a_{\min} < a_{\max} < 1$.

Envelope, admissibility, and local geometry. The envelope maximizes over centers. A fixed region can lose its low branch at a different temperature, and restricting θ to $(a_{\min}, a_{\max})$ can exclude the maximizing center. The recovered center must therefore be checked before treating an envelope as an attainable boundary. At a transverse fold with $H = H_a = 0$, $H_{aa} \neq 0$, and $H_T \neq 0$, Taylor expansion gives

$$a_\pm - a_* = \pm\sqrt{-\frac{2H_T(a_*, T_*)}{H_{aa}(a_*, T_*)}(T - T_*)} + o(\sqrt{|T - T_*|}).$$

The relaxation time scales as the inverse square root of distance to this fold. At a half-response tangency, H_T can vanish; the displayed scaling then requires another unfolding. The exact claims here are the locus and the envelope; we make no claim about the global cusp geometry.

E.4 How many equilibria can a single link produce?

For the log, shifted-log, and probability-logit links, the Schwarzian derivative $\mathcal{S}(f) = f'''/f' - \frac{3}{2}(f''/f')^2$ satisfies $\mathcal{S}(\ell) = \frac{1}{2}(\ell')^2$. The logistic has Schwarzian $-1/2$, so the composition identity gives

$$\mathcal{S}(\rho_T) = \tfrac{1}{2}(1 - T^{-2})(\ell')^2.$$

The map $a^*(u)$ is Möbius, hence $\mathcal{S}(\Phi) = \mathcal{S}(\rho_T)$ for $\Phi = a^* \circ \rho_T$. When $T < 1$, the function $v = (\Phi')^{-1/2}$ is strictly convex since $v'' = -\frac{1}{2}\mathcal{S}(\Phi)v > 0$. Four fixed points of Φ would give, by Rolle's theorem, three points with $\Phi' = 1$, hence $v = 1$, contradicting strict convexity. At $T = 1$, Φ is a nonidentity Möbius map and has at most two fixed points; it cannot be the identity because its upper image is strictly below $a_{\max}$. When $T > 1$, the inverse map has negative Schwarzian on its image and the same fixed points, giving the same three-root bound. Boundary signs then permit at most two attracting equilibria. This argument applies to the specified single-link families. An empirical mixture of many thresholds can have more roots.

E.5 A FINITE-SAMPLE RESPONSE CERTIFICATE

Let $\eta_1, \dots, \eta_n$ be independent draws from a fixed threshold distribution and define the raw-accuracy-scale response

$$\hat{\rho}_T(a) = \frac{1}{n}\sum_i \sigma((a-\eta_i)/T), \qquad \rho_T(a) = \mathbb{E}[\sigma((a-\eta)/T)].$$

With probability at least $1-\alpha$,

$$\sup_a \varphi\rho_T'(a)(a_{\max}-a) \le \sup_a \varphi\hat{\rho}_T'(a)(a_{\max}-a) + \frac{\varphi(a_{\max}-a_{\min})}{2T}\sqrt{\frac{\log(2/\alpha)}{2n}}. \tag{13}$$

Indeed, the Dvoretzky–Kiefer–Wolfowitz inequality bounds the empirical CDF error by the square-root term. Integration by parts converts that error into a derivative error multiplied by the total variation of the logistic density, $1/(2T)$. Applying Proposition 3 to a right-hand side at most one gives a population uniqueness certificate. The $1/T$ factor explains why sharp responses need more threshold samples. These raw-scale temperatures differ from the link-scale temperatures in Eq. (12).

F RECOVERY TIME, BUDGET, AND REACHABILITY

Assume $a_{\min} < \theta < a_{\text{prac}}$, $a_0 < \theta$, and $\delta > \delta_{\text{crit}}$. Let $A_\delta = a^*(\delta)$ and $r_\delta = \mu(1+\varphi\delta)$. The crossing time from Appendix A is

$$t_\delta = \frac{1}{r_\delta}\log\frac{A_\delta - a_0}{A_\delta - \theta}.$$

It decreases strictly with δ: the rate increases and the ratio $1+(\theta-a_0)/(A_\delta-\theta)$ decreases. Moreover,

$$A_\delta - \theta = \frac{\varphi(a_{\max}-\theta)}{1+\varphi\delta}(\delta - \delta_{\text{crit}}),$$

so crossing time diverges logarithmically as the floor approaches its threshold from above. A feasible recovery condition is therefore distinct from a timely recovery budget.

F.1 TRAINING VOLUME AND IMMEDIATE UTILITY

Let $p > 0$ denote incoming regional queries per unit time; when interpreting κ as practice per case, write $\kappa = \alpha p$. While the floor binds, forced training volume is

$$V_\delta = \delta p t_\delta.$$

At a positive per-case acquisition or training price d, spend is dV_δ. Both volume and this spend diverge as $\delta \downarrow \delta_{\text{crit}}$. The price d is separate from the fallback penalty λ in the routing objective.

Against a baseline that remains fully deferred, the floored policy's immediate routing-utility shortfall before crossing is

$$R_\delta = \delta p \int_0^{t_\delta} (\theta - a(t))\,dt = \delta p\left[\frac{\theta-a_0}{r_\delta} - (A_\delta-\theta)t_\delta\right]. \tag{14}$$

The second term vanishes at the critical limit, leaving a finite positive limit. Utility regret and training spend therefore measure different costs: a long interval of nearly competitive student answers can accumulate little regret while still consuming many training opportunities.

After crossing, competence follows $a(t_\delta + s) = a_{\text{prac}} - (a_{\text{prac}} - \theta)e^{-r_1 s}$, where $r_1 = \mu(1+\varphi)$. Its cumulative routing-utility gain over the deferred baseline pays back Eq. (14) at the unique $\Delta > 0$ satisfying

$$(a_{\text{prac}} - \theta)\left[\Delta - \frac{1-e^{-r_1\Delta}}{r_1}\right] = R_\delta/p.$$

Thus the payback horizon is $t_\delta + \Delta$, with $\Delta \le R_\delta/[p(a_{\text{prac}}-\theta)] + 1/r_1$. It inherits the logarithmic crossing-time divergence. This calculation fixes the comparator and includes the routing penalty; pricing additional labels and parameter updates requires their own costs. It is not a payback estimate for the language-model experiment.

F.2 Margins under uncertainty

Suppose a validated one-sided model-error bound gives, below θ,

$$\dot{a} \geq \mu\{\varphi\delta(a_{\max} - a) - (a - a_{\min}) - \epsilon\}, \qquad \epsilon \geq 0.$$

The lower affine drift is positive throughout $[a_{\min}, \theta]$ if

$$\delta > \frac{\theta - a_{\min} + \epsilon}{\varphi(a_{\max} - \theta)}.$$

Scalar comparison then bounds the true crossing time by the same logarithmic formula with A_δ replaced by $(\varphi\delta a_{\max} + a_{\min} - \epsilon)/(1 + \varphi\delta)$. This requires well-posed forward dynamics and a feasible right-hand side below one. It translates a bounded drift error into extra exposure margin.

Alternatively, write $\tau = (\theta - a_{\min})/(a_{\max} - \theta)$ and suppose valid simultaneous intervals give $\tau \leq \tau_U$ and $0 < \varphi_L \leq \varphi$. A requested margin-$m$ floor is

$$\delta_{\text{request}} = (1 + m)\tau_U/\varphi_L, \qquad m > 0.$$

On the coverage event it exceeds the true threshold with the requested margin. If the request exceeds one, capping it does not certify that margin; crossing under the maximal floor still requires $\delta_{\text{crit}} < 1$. Static accuracy audits alone do not supply intervals for the dynamical parameters. One identification route fits relaxation under two positive exposure levels: if μ and $a_{\min}$ are known, each equilibrium obeys $(a^*(u) - a_{\min})/u = \varphi(a_{\max} - a^*(u))$, so two distinct equilibria identify the slope and intercept. Valid uncertainty must account for the fitted dynamics and temporal dependence.

F.3 Recovery with a learned response

For a continuous response ρ and target $s < a_{\max}$, define

$$q(a) = \frac{a - a_{\min}}{\varphi(a_{\max} - a)}, \qquad D_s = \{a \in [a_{\min}, s] : \rho(a) \leq q(a)\},$$

$$\delta_\rho(s) = \begin{cases} \max_{a \in D_s} q(a), & D_s \neq \emptyset, \\ 0, & D_s = \emptyset. \end{cases}$$

If $\delta_\rho(s) < \delta \leq 1$, the clipped response $\max\{\rho, \delta\}$ gives positive drift throughout $[a_{\min}, s]$. On D_s, the floor exceeds the exposure needed for positive drift; outside it the base response already does so. Compactness gives a minimum drift $v_{\min} > 0$ and a crossing-time bound $(s - a_0)/v_{\min}$. This supplies a response-specific sufficient floor. Substituting an effective jump into the sharp formula is a comparator, not a general bound for arbitrary learned responses.

Stability to response error. Suppose $H'_\rho \leq -m < 0$, $\|\hat{\rho} - \rho\|_\infty \leq \epsilon_0$, and $\|\hat{\rho}' - \rho'\|_\infty \leq \epsilon_1$. If $\varphi[(a_{\max} - a_{\min})\epsilon_1 + \epsilon_0] < m$, then $H'_{\hat{\rho}} < 0$, so the learned response retains a unique equilibrium. The equilibrium shift is bounded by

$$|\hat{a}^* - a^*| \leq \frac{\varphi(a_{\max} - a_{\min})\epsilon_0}{m}.$$

This follows by bounding drift and derivative differences and applying the mean value theorem. Near a fold, m becomes small, making the response-error margin consequential.

F.4 When an additional training intervention is needed

Replace the practice term by $\kappa u(a)g(a)(a_{\max} - a)$, where g is nonnegative, continuous, and locally Lipschitz, and is positive above $a_{\min}$. If

$$\sup_{a \in (a_{\min}, \theta)} \frac{a - a_{\min}}{\varphi g(a)(a_{\max} - a)} > 1,$$

then at some state below θ even full exposure has negative drift. If the full-exposure drift is positive at a lower state, continuity gives an intervening zero that solutions cannot cross upward; if it is already nonpositive throughout the lower region, recovery from there is likewise blocked. More directly, any state where maximal drift is strictly negative supplies a barrier to upward passage from below under every $u \leq 1$. A sufficiently strong floor can therefore preserve a skill before it enters a low-return region, while subsequent recovery can require a change in the training process. Teacher feedback is one such change experimentally, but its effect is not derived from this modified scalar law.

G ALLOCATING EXPOSURE ACROSS SKILLS

G.1 WHY THE FLOOR MUST REACH EACH SKILL

Let subtypes k have masses $w_k > 0$, with $\sum_k w_k = 1$, and independent regional dynamics. An aggregate quota $\sum_k w_k u_k \geq \delta$ permits $u_{k'} = 0$ whenever the other subtypes have total mass at least δ. In that allocation, competence in k' decays toward its residual floor even if it could sustain a practiced state under full exposure. Conversely, under sharp routing, subtype floors $\delta_k > \delta_{\mathrm{crit},k}$ restore the scalar recovery guarantee separately for every feasible subtype. The result is about exposure at the granularity at which learning occurs. It does not guarantee that observable strata coincide with every latent skill.

G.2 SHARED PARAMETERS AND RETAINED RISK

An exposure floor also has a population interpretation without a scalar competence state. Let $R_c(w) \geq 0$ be stratum-specific risk, $q_c > 0$ its population share, and $u_c \in [0, 1]$ its exposure weight. Assume the minimizers below exist. If every $u_c \geq \delta > 0$ and $w_u \in \arg\min_w R_u(w)$ for $R_u(w) = \sum_c u_c q_c R_c(w)$, then for any full-data minimizer w_{full},

$$R_c(w_u) \leq \frac{R_u(w_u)}{\delta q_c} \leq \frac{R_u(w_{\mathrm{full}})}{\delta q_c} \leq \frac{\max_{c'} R_{c'}(w_{\mathrm{full}})}{\delta q_c}. \tag{15}$$

The first inequality uses nonnegativity, the second uses optimality, and the third uses $\sum_c u_c q_c \leq 1$. Positive exposure prevents a stratum's loss from disappearing from the objective. The bound can be loose for rare strata and need not improve on the trivial bound for bounded losses. It concerns population minimization; it does not establish finite-sample generalization or convergence of LoRA updates.

For a fixed predictor and a finite stratum of N bounded-loss cases, uniform sampling of M cases without replacement gives $\Pr(|\hat{R} - R| > \epsilon) \leq 2e^{-2M\epsilon^2}$; a finite-population correction can sharpen it (Serfling, 1974). Independent Bernoulli sampling and fixed-quota sampling have different count distributions, even when both protect exposure. Adaptive-round guarantees require conditional bounds and an appropriate union or martingale argument. Generalization for a fitted predictor additionally needs a complexity or stability assumption.

G.3 PRACTICE TRANSFER

Let a nonnegative row-stochastic matrix K distribute practice across regions, with incoming rates p_j, practice coefficients α_i, and decay rates μ_i. Consider

$$\dot{a}_i = \alpha_i \Big(\sum_j K_{ij} p_j \rho_j(a_j) \Big) (a_{\max,i} - a_i) - \mu_i (a_i - a_{\min,i}).$$

Practice from source j acts on the *destination's* remaining competence gap. Define $\psi_{ij} = \alpha_i p_j K_{ij} / \mu_i$ and $L_j = \sup |\rho_j'|$. The equilibrium response map has componentwise Jacobian bound

$$B_{ij} = (a_{\max,i} - a_{\min,i}) \psi_{ij} L_j.$$

Thus $\max_i \sum_j B_{ij} < 1$ gives a contraction in the sup norm. More generally, if the spectral radius of B is below one, choose $v \gg 0$ and $q < 1$ with $Bv \leq qv$; the map is a contraction in $\|x\|_{\infty,v} = \max_i |x_i| / v_i$. These are uniqueness and convergence statements for full-relaxation iteration.

Under sharp routing with floors, region i receives enough practice to cross its threshold using floor exposure alone if

$$\sum_j \psi_{ij} \delta_j > \tau_i := \frac{\theta_i - a_{\min,i}}{a_{\max,i} - \theta_i}. \tag{16}$$

The proof compares its drift to the affine flow driven by the guaranteed constant exposure on the left. That lower flow has equilibrium above θ_i, hence crosses in finite time. The inequality is a worst-case guarantee: an already-active neighbor can supply more than its floor and enable earlier recovery. For $K = I$ it reduces to Eq. (6). Source-indexed practice/decay ratios are valid only when those rates are homogeneous; heterogeneous rates require the destination-indexed α_i / μ_i above.

H MONITORING RETAINED COMPETENCE

Floors can supply evaluation evidence as well as practice. For a fixed checkpoint, let $Z_i \in \{0, 1\}$ be student correctness on incoming case i, observed when $K_i = 1$, with known inclusion probability $\pi_i \geq \delta > 0$. If the draw does not change that case's correctness, the Horvitz–Thompson mean

$$\hat{a}_{\mathrm{HT}} = \frac{1}{n} \sum_{i=1}^{n} \frac{K_i Z_i}{\pi_i}$$

is unbiased for the case-population mean because $\mathbb{E}[K_i Z_i/\pi_i \mid Z_i, \pi_i] = Z_i$ (Horvitz & Thompson, 1952). For independent draws, boundedness in $[0, 1/\delta]$ gives the conservative deviation radius

$$\frac{1}{\delta}\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

The self-normalized Hájek ratio $\sum_i K_i Z_i/\pi_i \,/\, \sum_i K_i/\pi_i$ is consistent under standard sampling conditions but is not exactly unbiased. Doubly robust estimators additionally use an outcome model (Dudík et al., 2011). Labels must eventually be available: positive routing probabilities alone do not reveal correctness.

With adaptive logging, propensities must be predictable from the prior history and the current case, before the routing draw. A pooled estimate then concerns the average correctness of the sequence of deployed checkpoints; it does not identify a different training history or the current checkpoint's accuracy without further assumptions. The main language-model experiments instead evaluate each checkpoint on a separate held-out set.

An aggregate quota need not give every case positive inclusion probability. With an all-or-none stratum rule, uniform top-up of a deferred stratum does. With a heterogeneous base router, an independent Bernoulli audit draw or explicitly computed inclusion probabilities are needed for the stated estimator. The stratum-level volume condition and the case-level logging condition serve different purposes.

Monitoring when the serving threshold is unreachable. When $a_{\mathrm{prac}} < \theta$, a floor can maintain competence and observability even though full exposure does not make keeping competitive. At equilibrium its recurring routing-utility price is $P(\delta) = \delta p[\theta - a^*(\delta)]$. On a chosen interval $[\delta_{\min}, 1]$, a monitoring design can minimize

$$J(\delta) = P(\delta) + \frac{\lambda_v}{4\delta^2 n}, \qquad \delta_{\min} > 0.$$

The second term is a conservative range-based variance proxy, not an exact HT variance. Continuity gives a minimizer; the smallest and largest minimizing floors are nondecreasing with audit value λ_v and nonincreasing with traffic n, by decreasing and increasing differences respectively (Topkis, 1998). This monitoring objective is separate from selecting a floor to cross a reachable threshold.

I RESPONSE-SHAPE EXPERIMENTS AND NUMERICAL CHECKS

The supplementary studies separate properties that are coupled in the language-model cascade. They test the scalar formulas, import empirical threshold shapes while controlling the learning parameters, and learn a routing response from a calibrated surrogate. We label them E1–E7. They complement the language-model experiments by isolating the conditions and design choices that govern the feedback.

I.1 E1: CHECKING THE ANALYTICAL BOUNDARIES

The verification uses $a_{\max} = 0.95$, $a_{\min} = 0.55$, and $\mu = 0.08$, with φ spanning 0.8–8. An independent damped-Newton solver solves the raw tangency equations $H = H' = 0$ with a numerical Jacobian and residual tolerance 10^{-12}. It does not use Eq. (11) to locate the tangency. Median relative log-link discrepancy is 3.5×10^{-15}, with maximum 4.3×10^{-13}; the probability-logit and shifted-log discrepancies are at most 2.2×10^{-15} and 5.0×10^{-13}, respectively. A sweep over ceilings $\{0.90, 0.95, 0.99\}$ and residual floors $\{0.45, 0.55, 0.65\}$ checks the dependence on the state interval.

Brute-force root-count maps near the log-link tip at $\varphi = 4$ approach the envelope from below: the largest resolved bistable temperatures are 0.1115/0.1130/0.1136 on 40/80/160 grids, compared with the closed-form 0.1141. At $\varphi = 8$, the envelope's maximizing center is 0.9559, outside the interval ending at 0.95; the admissible boundary is 0.1217, below the unconstrained envelope 0.1240. This distinguishes a center-specific attainable transition from the envelope. Explicit-Euler integration at step 10^{-4} also checks the floor crossing-time formula, with median error 0.0004%. These numerical checks of the analytical results cover only the tested parameter values.

I.2 E2A: EMPIRICAL THRESHOLD SHAPES AT A COMMON CENTER

CIFAR-10H provides approximately 51 human judgments per CIFAR-10 test image (Peterson et al., 2019). Per-image agreement with the modal human label supplies a threshold, and within-class spread ranges from 0.053 to 0.122. The controlled comparison recenters every class to 0.80 and fixes $(\varphi, a_{\max}) = (4.875, 0.95)$ and $a_{\min} = 0.50$. Only the empirical spread and shape vary. The response averages raw-scale logistic thresholds at $T = 0.02$.

In this construction, the six lowest-spread classes have two attracting equilibria and the four highest-spread classes have a unique equilibrium. The split lies between spreads 0.083 and 0.093. This instantiates the response-concentration mechanism while holding the scalar learning parameters fixed. Multimodal thresholds can also create disjoint bistable temperature windows, illustrating why a single Gaussian approximation need not capture the full response.

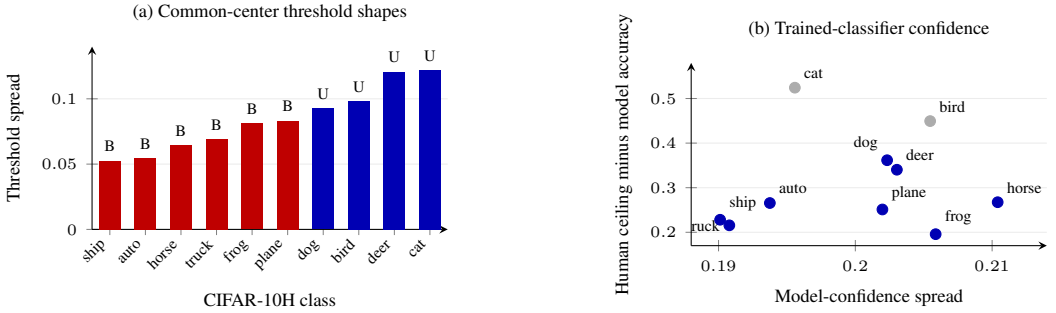

Figure 4: Threshold-shape studies (E2). Left: CIFAR-10H threshold spread at a common center and fixed learning parameters; B denotes bistability and U uniqueness. The descriptive boundary band is 0.088–0.106. Right: classifier confidence spread versus human ceiling minus model accuracy; blue denotes positive learning-ratio calibration, gray denotes separately analyzed classes. Derivative-certificate results are reported in the text.

Measurement and numerical status. The data provide agreement distributions, while the competence dynamics are imposed. Selecting and scoring the modal label on the same votes can affect the agreement estimate; we therefore use annotator half-splits that choose consensus on one half and score it on the other. The recentered partition is then rederived. The descriptive 5th–95th percentile band for the algorithm-defined boundary midpoint is $(0.088, 0.106)$ under image bootstrap resampling; it is not a confidence interval for a universal critical dispersion. Without recentering, every class is unique at $T = 0.02$, although much sharper raw mixtures can have multiple roots, including more than three. The three-root theorem for a single link does not apply to arbitrary mixtures.

Root counts use interval enclosures for H, H' under a stated binary64 forward-error model, with high-precision oracle regression. Counts are returned only when the unresolved set is empty. These are interval-validated counts conditional on that arithmetic model, rather than outward-rounded computer-assisted proofs. Mean thresholds alone do not justify scalar support assumptions; the isolator uses the individual threshold atoms.

I.3 E2B: CONFIDENCE DISTRIBUTIONS FROM A TRAINED CLASSIFIER

A CIFAR-10 classifier uses whitened random-patch features and a multinomial softmax head trained on 50,000 images; its test accuracy is 64.4%. Per-case maximum softmax confidence supplies the thresholds. The construction calibrates $\varphi_c = 2.2(\mathrm{acc}_c - a_{\min})/(a_{\max,c} - \mathrm{acc}_c)$ for the eight classes

on which this quantity is positive. The other classes are analyzed directly over a 21-point sweep of the observed positive-φ range.

At the operating scales, the threshold dispersion supports a unique equilibrium. Seven calibrated classes pass the sufficient derivative bound directly; the remaining class is resolved by interval-validated root counting. All calibrated classes remain unique for the evaluated $T \geq 10^{-3}$. At substantially sharper scales, near-tangent triples appear in individual finite mixtures. Fitting temperature scaling on even-index examples and evaluating on the disjoint odd-index half changes ECE from 0.117 to 0.023. The uniqueness conclusion is preserved, while fewer classes pass the sufficient derivative certificate. This separates the actual equilibrium structure from the strength of one sufficient bound.

I.4 E2c: Learning the response with a surrogate head

The learned-head study uses a two-dimensional, three-class synthetic distribution with known posteriors. A hard component of mass 0.75 smoothly mixes two classes, with $\eta_{\max}(x) = 0.5 + 0.22|\tanh(x_1/0.3)|$; much of this component has thresholds in $[0.71, 0.72]$. An easy component of mass 0.25 has near-unit maximum posterior. At each competence level in $[0.55, 0.95]$, a $K + 1$ head on 500 random Fourier features is trained with plug-in expert weight $(\hat{a} - 0.10)_+$. Expert-correctness draws estimate $\hat{a}$, and routing is measured on 60,000 held-out points; the response summary aggregates the fitted heads.

The learned response has mean absolute deviation 0.068 from the calibrated response; its maximum deviation away from the concentrated band is 0.081. At the band it can reach 0.188, where a small posterior error moves a large fraction of the routing mass. Driving the scalar flow with this learned response gives a unique equilibrium at $\varphi = 1.2$ and two attracting equilibria at $\varphi = 5$. In the latter setting, the low state is 0.561 and the practiced state is 0.840. A floor near 0.250 restores the practiced state, agreeing to sweep resolution with the learned-response threshold 0.246 from Appendix F. The sharp effective-threshold comparator is 0.322.

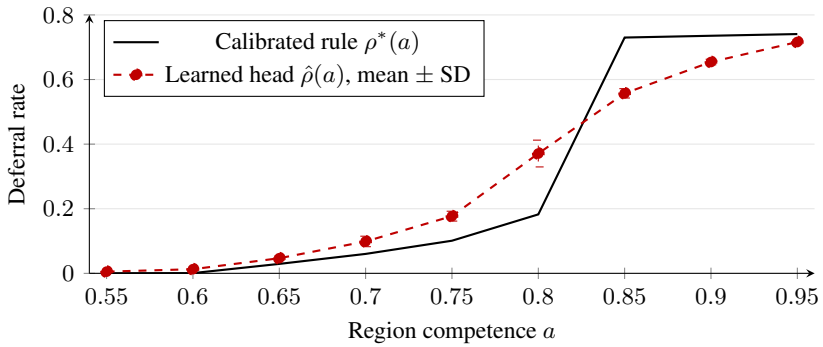

Figure 5: A trained surrogate head supplies a competence-dependent exposure response (E2c). The curve shows the mean with standard-deviation bars against the calibrated rule; its concentrated rise supplies the feedback studied in the equilibrium and floor calculations.

The loop and floor sweep use the same interpolated learned response, so the threshold agreement is partly internal. This experiment establishes that surrogate training can instantiate the response shape; the language-model experiment separately tests changes in the learner's actual parameters and held-out performance.

J Memory and exposure in a stationary cascade

J.1 E3 protocol and operating points

The stationary `digits` cascade isolates the role of routed training data with a small convex learner. Disjoint streams contain 300 bootstrap, 650 feedback, 350 calibration, and 497 evaluation examples. A fixed logistic fallback uses raw pixels; the student uses six PCA components and is refit on a four-round window. Each of 50 rounds draws 90 feedback examples without replacement from the

feedback pool, reused across rounds. Predicted student classes define observable strata. A stratum is kept when calibration student precision is at least fallback accuracy on that stratum minus $\lambda = 0.15$. True classes are used for reporting only. The same deployed policy generates both feedback exposure and evaluation routing.

Table 11 reports the complete policy comparison. A 15% stratum floor changes student accuracy from 0.398 ± 0.034 to 0.826 ± 0.009, while reducing fallback use from 0.971 ± 0.029 to 0.276 ± 0.037. Cumulative replay yields student accuracy 0.854 ± 0.006 and utility 0.859 ± 0.006. These outcomes show two effective ways to maintain training access: reserve new examples, or retain the examples already acquired when memory permits.

Table 11: Stationary cascade operating points (E3), aggregate final-five-round means ± SE. These descriptive estimates are exploratory. Deferral denotes the fallback-call fraction; utility is accuracy minus 0.15 times that fraction.

Policy	System acc.	Deferral	Utility	Student acc.
naive greedy	0.935 ± 0.008	0.971 ± 0.029	0.790 ± 0.005	0.398 ± 0.034
exposure floor 15%	0.881 ± 0.006	0.276 ± 0.037	0.840 ± 0.006	0.826 ± 0.009
count-targeted random	0.880 ± 0.005	0.283 ± 0.023	0.838 ± 0.005	0.819 ± 0.008
UCB (pure)	0.843 ± 0.007	0.007 ± 0.007	0.842 ± 0.007	0.843 ± 0.006
UCB + floor	0.842 ± 0.007	0.004 ± 0.004	0.841 ± 0.007	0.842 ± 0.006
distillation replay	0.889 ± 0.007	0.302 ± 0.045	0.844 ± 0.006	0.823 ± 0.010
floor + distillation	0.878 ± 0.006	0.226 ± 0.032	0.844 ± 0.006	0.834 ± 0.007
periodic refresh	0.917 ± 0.011	0.692 ± 0.102	0.813 ± 0.008	0.586 ± 0.054
frozen student	0.886 ± 0.008	0.237 ± 0.043	0.850 ± 0.005	0.845 ± 0.005
cumulative history	0.894 ± 0.007	0.228 ± 0.048	0.859 ± 0.006	0.854 ± 0.006
class-balanced replay	0.897 ± 0.007	0.242 ± 0.046	0.860 ± 0.005	0.854 ± 0.006
floor + cumulative	0.886 ± 0.006	0.163 ± 0.029	0.862 ± 0.006	0.862 ± 0.006

What the comparisons identify. The floor improves routing utility at the specified penalty by exchanging some system accuracy for a large reduction in fallback use. It does not Pareto-dominate greedy routing. The replay family attains the highest utilities in this stationary setting. Memory-window ablations show stronger loss with shorter memory; exposure protection is useful when replay is restricted. Rerouting both policies at $\lambda = 0.05$ favors greedy utility, whereas a $\lambda = 0.30$ probe reverses it. The utility ranking therefore depends on the operating price and the induced routing policy, not only on the learner's retained accuracy.

Count-targeted random exposure uses the paired floor's per-round forced training count as a target, truncated when the rejected pool is smaller. It does not match exposure by stratum or at evaluation. The selected class has final accuracy 0.803 under the floor versus 0.780 under count-targeted random exposure. The selected-class comparison is post hoc; it does not establish general worst-group superiority. Distillation uses 60 fallback pseudo-labels per round, and periodic refresh uses 80 oracle labels every ten rounds. Their auxiliary calls are counted but excluded from the table's routing utility. Thus improvements in this utility are not complete training-cost comparisons.

Interpretation of the summaries. The tables report aggregate endpoints, with standard errors describing algorithmic variation on the benchmark. The analysis is exploratory, including its intervals, because evaluation continued after an interim inspection without a prespecified stopping rule. The operating points compare retained competence, serving accuracy, and fallback use at the stated routing price.

J.2 E4: ADAPTING EXPOSURE TO CALIBRATION GAPS

The adaptive-floor study uses the same data and deployed-policy discipline. It raises the floor to 0.25 when calibration student accuracy trails the fallback and otherwise uses 0.05, compared with fixed floor 15%. The adaptive and fixed utilities are 0.839 ± 0.005 and 0.840 ± 0.006, respectively; their paired difference is -0.001 $[-0.006, +0.005]$. Both preserve a useful student compared with greedy routing. The adaptive rule forces more learning-side exposure in the lagging strata; its lower fallback use is not evidence of a smaller training budget. This is a gap-based heuristic, distinct from estimating the dynamical parameters needed for the certified margin rule in Appendix F.

K MONITORING, TRANSFER, AND ALLOCATION DIAGNOSTICS

K.1 E5: OBSERVING COMPETENCE THROUGH SELECTED TRAFFIC

The logging simulation has difficulty $s \sim \mathcal{U}[0,1]$ and correctness probability $a(s) = 0.85 - 0.30s$, giving population accuracy 0.70. The logging probability is the logistic $\sigma((s-0.5)/0.15)$ clipped below at δ. At $\delta = 0.2$ and $n = 2000$, aggregate simulation bias is -0.047 for the unweighted selected-traffic mean, -0.000 for HT, and -0.000 for Hájek. The corresponding nominal 90% conservative HT interval has empirical coverage 1.00. The comparison demonstrates the distinction between measuring the cases selected by a policy and measuring the underlying case population.

A separate simulation evaluates Eq. (13) on Gaussian threshold pools at $T = 0.1$ and $\varphi = 1.8$. Increasing sample size or dispersion makes the sufficient bound more informative. Pools near the boundary can remain uncertified. Trials on directly verified bistable pools record no false certifications; the formal guarantee remains the stated confidence level. These simulations concern static response identification, not estimation of the full evolving learning process.

K.2 E6: LEARNING TRANSFERRED BETWEEN REGIONS

Three-region simulations test the network conditions in Appendix G. The spectral-radius condition certifies uniqueness for instances that the row-sum bound leaves unresolved. Certified systems converge to a common equilibrium from the tested initial states, supporting the analytical uniqueness conditions.

In a two-region example, region 2 receives a direct floor 0.15, below its isolated threshold 0.370. Its isolated final competence is 0.642, while transfer from a protected neighbor raises it to 0.817. When the neighbor contributes only its floor, the rescue boundary agrees with the worst-case inequality in Eq. (16) to grid resolution. When that neighbor itself reaches a practiced state, its extra exposure supports an earlier cascade of recovery. The modeled coupling allows practice in a protected neighbor to complement direct exposure.

K.3 E7: PROTECTING A SKILL WITHIN A LARGER REGION

The allocation diagnostic contains two equally weighted subtypes, with common constants

$$(a_{\min}, a_{\max}, \mu, p) = (0.55, 0.95, 0.08, 0.5).$$

The easy subtype has threshold 0.60, $\varphi = 3$, and starts practiced. The harder subtype has threshold 0.78, $\varphi = 6.0$, and starts below threshold; its sharp-response critical floor is 0.225. A router that minimizes immediate regret allocates forced exposure first where the adaptive decision-maker's margin is greatest.

An aggregate quota of 0.3 is met by the practiced subtype alone, leaving the harder subtype at 0.550. Applying that same floor to each subtype brings the harder one to 0.891. Even a binding aggregate quota of 0.55 leaves it at 0.700 because its residual exposure is only 0.100, below threshold. An equal-realized-budget aggregate control also leaves it near the low state, at 0.554. The key variable is where practice goes. In the sharp-response control, observed crossing time 6.12 agrees with the formula's 6.13; the smoothed response at $T = 0.04$ also recovers under subtype protection.

Reproducibility and study boundaries. Scripts for E1–E7, numerical verification, and a regression test suite are released with the code; the studies reproduce in roughly an hour of CPU time. The image-cascade evidence is in Appendix L.

The combined evidence separates four levels: exact statements about the specified scalar or finite-region model; numerical checks of those statements; controlled response-shape and stationary-learning studies; and finite-horizon parameter-update experiments. The mathematical conditions explain how the mechanism can arise. The deployment interventions establish its consequences in the evaluated learners and show concrete ways to maintain access to learning.

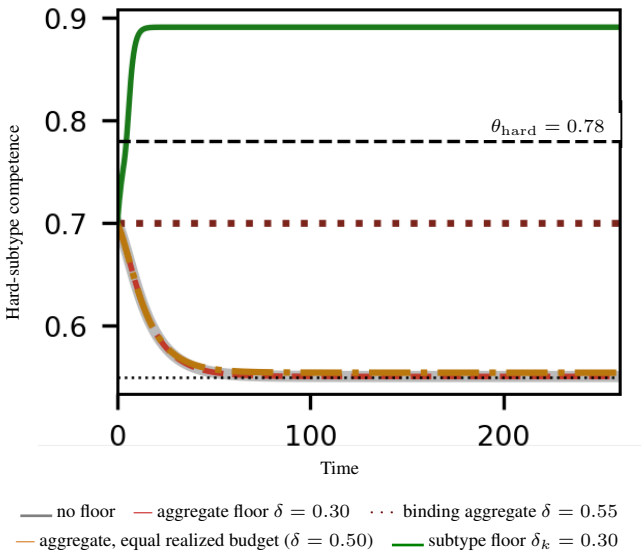

Figure 6: Subtype protection in the deterministic allocation diagnostic (E7). The aggregate quota can leave the harder subtype below its recovery threshold even when it requires real expenditure. Protecting exposure per subtype restores the practiced state. Budget comparisons are specific to the simulated horizon.

L IMAGE CASCADES

The routing penalty is $\lambda = 0.15$, with a 25% exposure floor and a 40-round horizon. Student logistic heads use Gaussian projections of dimension 16 for CRC100K and 32 for CIFAR-100, regularization $C = 1$, and at most 1,500 solver iterations. Each round supplies 1,500 feedback examples for CRC100K or 1,000 for CIFAR-100. The student is refit from scratch on a bounded memory window, so continued exposure determines which strata remain represented in training. These experiments test exposure protection and replay under this learning procedure.

L.1 PROTOCOL AND EXPOSURE INTERVENTIONS UNDER STALE CALIBRATION

Dataset	Policy	System acc.	Deferral	Utility	ΔU	Worst class
CRC100K	Greedy	.9104	.358	.8567	—	.347
	Floor 25%	.9071	.253	.8691	+.0124	.762
	Floor 25% + replay	.9083	.237	.8727	+.0160	.805
CIFAR-100	Greedy	.8184	.982	.6710	—	.000
	Floor 25%	.8574	.750	.7450	+.0740	.740
	Floor 25% + replay	.8617	.750	.7493	+.0782	.945

Table 12: **Exposure interventions in image cascades under shift with stale calibration.** Aggregate terminal values. Worst class is the student's minimum per-class accuracy. Greedy and floor-only students use a four-round memory window; replay keeps all retained examples.

CRC100K uses 100,000 colorectal patches in nine classes (Kather et al., 2018) with frozen CONCH features from the TITAN release (Ding et al., 2025); CIFAR-100 uses its 50,000-image training split with frozen CLIP ViT-L/14 features (Krizhevsky, 2009; Radford et al., 2021). The pool is divided into disjoint fallback-training, bootstrap, calibration, evaluation, and feedback streams, and all policies receive the same per-round feedback sample. The fallback, region assigner, and student are multinomial logistic heads: the fallback and assigner use full features, the student a fixed low-dimensional random projection, refit from scratch each round on retained data. Strata are predicted classes; a stratum is kept when student calibration accuracy is at least fallback accuracy minus a penalty λ. At round $r \in \{0, \ldots, 39\}$, class-c features shift by $\frac{r}{39}(m_c^{\text{target}} - m_c^{\text{source}})$; for CRC the centroids come from stain-normalized and unnormalized versions of the data, and for CIFAR the displacement is a class-specific Gaussian vector scaled by the within-class standard deviation. Calibration is unshifted in Table 12, so estimated regional risks become stale. Utility is $U = A - \lambda C$

with A macro-averaged system accuracy and C the deferral fraction. The CIFAR-100 displacement grows class separability over the horizon, which is why a low-dimensional student can exceed the accuracy typical of static linear probes; CIFAR-100 worst-class values are therefore best read as comparisons between policies under the same shift.

L.2 Combining calibration refresh and exposure protection

To separate stale calibration from exposure deprivation, a CRC100K factorial crosses calibration (stale, or re-estimated every round on a fresh labelled draw from the current distribution), a 25% floor, and replay, with always-student, frozen-student, and oracle-routing references. The factorial uses CONCH v1 features, whereas Table 12 uses features from the TITAN release; it reproduces the signs and ordering with somewhat lower deferral and is not an exact feature-matched replication.

Calibration	Policy	Utility	Worst class	Net of labels
Stale	Greedy	$0.868 \pm .010$	$0.511 \pm .092$	0.868
Stale	Floor 25%	$0.884 \pm .009$	$0.828 \pm .026$	0.884
Refreshed, large budget	Greedy	$0.937 \pm .003$	$0.688 \pm .107$	0.337
Refreshed, large budget	Floor 25%	$0.935 \pm .004$	$0.828 \pm .031$	0.335
Refreshed, small budget	Greedy	$0.934 \pm .003$	$0.477 \pm .093$	0.904
Refreshed, small budget	Floor 25%	$0.934 \pm .004$	$0.841 \pm .025$	0.904
Refreshed, smallest budget	Greedy	$0.911 \pm .009$	$0.380 \pm .110$	0.896
Refreshed, smallest budget	Floor 25%	$0.930 \pm .005$	$0.841 \pm .026$	0.915
Refreshed, large budget	Frozen student	$0.849 \pm .009$	$0.543 \pm .037$	—
—	Always student (window)	$0.935 \pm .004$	$0.866 \pm .012$	—

Table 13: Mean $\pm$ SE over three seeds; net of labels prices refresh labels at λ. Calibration refresh and exposure floors improve different combinations of serving utility and worst-class retention across the evaluated label budgets. Terminal values.

At the large refresh budget, greedy utility rises from 0.868 to 0.937 and worst-class accuracy from 0.511 to 0.688 (Table 13). Adding a floor under refreshed calibration raises worst-class accuracy further to 0.828, with utility 0.935. The combination retains a worst-class benefit after routing estimates are refreshed.

At the small refresh budget, the floor raises worst-class accuracy from 0.477 to 0.841 at the same reported utility of 0.934; both policies have utility 0.904 after pricing refresh labels. At the smallest budget, it raises worst-class accuracy from 0.380 to 0.841, utility from 0.911 to 0.930, and utility net of labels from 0.896 to 0.915. Thus continued exposure supports retention across calibration budgets, with especially large worst-class gains when refresh labels are limited. The large-budget frozen student reaches worst-class accuracy 0.543, while the always-student window reference reaches 0.866. Together with the floor comparisons, these references show the value of continued learning with broad training coverage during the shift.

L.3 Refreshed calibration on CIFAR-100

The same factorial on CIFAR-100 uses the original CLIP ViT-L/14 features of Table 12. The repeated stale-calibration comparison gives floor-only worst-class accuracy 0.737 and floor-plus-replay accuracy 0.952 (Table 15), close to 0.740 and 0.945 in the original comparison. The refreshed arm re-estimates student and fallback accuracies every round on 3,000 fresh labelled examples from the current distribution; 150–600 labels give the sensitivity rows.

Under the ramp shift, refreshing with 3,000 labels per round raises greedy utility from 0.674 to 0.891, while worst-class accuracy remains 0.000 (Table 14). Combining refresh with a floor raises utility to 0.991 and worst-class accuracy to 0.928, matching the always-student window reference and reducing deferral to zero. The frozen student has worst-class accuracy 0.000 under the same refresh budget. These comparisons demonstrate the benefit of continued exposure alongside refreshed routing estimates.

The smaller refresh budgets produce different routing patterns: greedy deferral falls from 0.341 at 3,000 labels to 0.023 at 150 labels, and worst-class accuracy rises from 0.000 to 0.415. Floor policies at 150–600 labels reach worst-class accuracy 0.923–0.927. Across these budgets, an explicit floor

Calibration	Policy	Deferral	Utility	Worst class	Net of labels
Stale	Greedy	0.968	0.674 ± .003	0.000 ± .000	0.674
Stale	Floor 25%	0.745	0.746 ± .003	0.737 ± .025	0.746
Refreshed (3,000)	Greedy	0.341	0.891 ± .006	0.000 ± .000	0.441
Refreshed (3,000)	Floor 25%	0.000	0.991 ± .001	0.928 ± .005	0.541
Refreshed (600)	Greedy	0.152	0.949 ± .006	0.000 ± .000	0.859
Refreshed (300)	Greedy	0.058	0.975 ± .002	0.005 ± .003	0.930
Refreshed (150)	Greedy	0.023	0.984 ± .001	0.415 ± .128	0.962
Refreshed (3,000)	Frozen student	0.906	0.669 ± .003	0.000 ± .000	—
—	Always student (window)	0	0.991 ± .001	0.928 ± .005	—

Table 14: CIFAR-100 calibration factorial. Terminal values, mean ± SE over ten seeds; no replay; net of labels prices refresh labels at λ. Floor rows at 150–600 labels reach worst-class 0.923–0.927. Table 15 repeats the factorial under a separability-neutral shift.

		Deferral		Utility		Worst class	
Calibration	Policy	Ramp	Neutral	Ramp	Neutral	Ramp	Neutral
Stale	Greedy	0.968	0.968	0.674	0.674	0.000	0.000
Stale	Floor 25%	0.745	0.711	0.746	0.610	0.737	0.047
Refreshed (3,000)	Greedy	0.341	0.920	0.891	0.680	0.000	0.000
Refreshed (3,000)	Floor 25%	0.000	0.607	0.991	0.615	0.928	0.048
Refreshed (600)	Greedy	0.152	0.886	0.949	0.673	0.000	0.000
Refreshed (300)	Greedy	0.058	0.766	0.975	0.662	0.005	0.000
Refreshed (150)	Greedy	0.023	0.656	0.984	0.644	0.415	0.000
Refreshed (3,000)	Frozen student	0.906	0.924	0.669	0.668	0.000	0.000
—	Always student (window)	0	0	0.991	0.568	0.928	0.163
Stale	Floor 25% + replay	0.748	0.672	0.750	0.661	0.952	0.226
—	Always student + replay	0	0	0.995	0.650	0.952	0.241

Table 15: CIFAR-100 calibration factorial under the constructed ramp shift of Table 14 and under a norm-matched, separability-neutral shift that moves every class by the same vector. Terminal values, means over the same ten seeds (paired splits and feedback). The ramp raises the between- to within-class scatter ratio from 0.45 to 1.46 and lets a freshly trained student reach 0.991 balanced accuracy; the neutral shift holds the ratio at exactly 0.45 and that student at 0.545, while degrading the fallback by the same amount. Under the neutral shift the refreshed floor's worst-class gain is 0.048 [0.033, 0.064] and its utility change -0.065 [-0.068, -0.062]; refreshing adds 0.006 utility; with cumulative replay the stale floor gains 0.226 worst-class accuracy for a utility cost of 0.013. Neutral-shift SEs are at most .014 (worst class) and .004 (utility).

consistently provides high worst-class retention under the ramp shift. Net utility also reflects the price of refresh labels: the 3,000-label floor yields 0.541 after label costs, while stale calibration with a floor yields 0.746.

The ramp raises the between- to within-class scatter ratio from 0.45 to 1.46. To assess exposure protection at fixed separability, Table 15 repeats the factorial on the same ten seeds using a norm-matched common translation of every class, which preserves the scatter ratio at 0.45. Stale greedy routing has deferral 0.968 and worst-class accuracy 0.000 under both shifts. Under the neutral shift, the refreshed floor raises worst-class accuracy from 0.000 to 0.048, with paired 95% interval 0.033–0.064, while utility changes from 0.680 to 0.615. The utility gain from refreshing greedy routing is 0.006 under this shift, compared with 0.217 under the ramp. The 150-label greedy policy retains worst-class accuracy 0.000 under the neutral shift.

Replay provides a larger retention gain at this harder operating point. With stale calibration, the floor plus cumulative replay reaches worst-class accuracy 0.226 and utility 0.661, compared with 0.000 and 0.674 for greedy routing, and 0.047 and 0.610 for the floor alone. It comes within 0.015 of the always-student replay reference's worst-class accuracy of 0.241, with utility 0.011 higher and deferral 0.672. The always-student window reference reaches worst-class accuracy 0.163 and utility 0.568. These comparisons identify exposure plus replay as an effective combination for maintaining broad student competence while continuing to use the fallback. The neutral-shift outcomes fall below the preregistered target of a worst-class gain of at least 0.30 with no utility loss; the measured gains quantify the retention available at the reported utility tradeoffs.