# SAME WEIGHTS, DIFFERENT FUTURES: STATE IDENTIFIABILITY IN ON-POLICY DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy distillation changes not only a student's predictions but also the states on which it receives future supervision. Thus, two students that currently appear identical may still have different probabilities of recovering a missing behavior within a finite query budget. Specifically, we study which representation of training state determines this recoverability. To formalize this, we define budgeted recoverability on a transition-sufficient state and show that an observable summary is sufficient exactly when all full states sharing that summary have the same recovery probability. The resulting observation-fiber diameter gives a sharp minimax lower bound on prediction error from the summary alone. We then construct two positive-probability histories of one finite OPD process that have identical observations but different recovery laws, and show that Adam can preserve such differences in optimizer state even when weights are identical. We further characterize scalar recovery regimes and decompose the value of finite-query intervention into state access and the signed reliability of teacher feedback. Empirically, common-future Transformer experiments realize the exact-weight separation, while optimizer cross-swaps test the same mechanism under ordinary training in a procedural task and Qwen3.5 reasoning on GSM8K. Prospective validation indicates that recovery-gap magnitude transfers more reliably than state ordering. Together, these results show that recoverability is determined by the state consumed by future transitions, not by current weights or predictions alone.

## 1 INTRODUCTION

On-policy distillation trains a student on prefixes or sequences generated by its own behavior and queries the teacher on those states. This coupling can reduce the mismatch between fixed supervised data and the distribution encountered at inference time, extending the classical view of distillation as predictive-distribution transfer (Hinton et al., 2015; Kim & Rush, 2016; Rusu et al., 2015). Yet it also changes what information is needed to predict future learning. In scheduled sampling and interactive imitation, learner behavior determines which examples or expert labels become available (Bengio et al., 2015; Huszár, 2015; Ross et al., 2011); modern on-policy language-model distillation makes this dependence explicit through student-generated sequences and occupancy-aware objectives (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; Wen et al., 2023). As a result, agreement in current predictions need not determine whether a missing behavior can be recovered within a finite query budget. Two runs can make identical predictions while differing in optimizer moments, replay contents, teacher state, random state, or update geometry. This raises the central question: *which representation of training state determines future recoverability?*

Three challenges make this question nontrivial. **First**, current logits and divergence can agree even when variables consumed by future updates differ. Experience replay and target networks already show how algorithmic memory can affect future behavior beyond current predictions (Schaul et al., 2015; Mnih et al., 2013). **Second**, student-induced occupancy couples access to learning: when a behavior becomes unlikely, the teacher states needed to recover it may also become unlikely to be visited. Interactive learning therefore emphasizes querying experts on learner-visited states (Rajaraman et al., 2021; Hoque et al., 2022), but access alone is insufficient because feedback can help or harm depending on its conditional reliability. **Third**, recoverability can vary across histories and query settings. A stable average recovery gap need not imply a stable ordering of individual states, and additional access need not have a uniform sign. In conclusion, these challenges call for an



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.

event-specific notion of information sufficiency together with separate analyses of state visitation and feedback reliability.

To answer this question, we define budgeted recoverability on a transition-sufficient training state under a declared query policy and transition kernel. Specifically, a declared observation map partitions full states into fibers that are observationally indistinguishable. Recoverability factors through the observation exactly when it is constant on every fiber, and any positive fiber diameter yields a sharp minimax lower bound on summary-only prediction error. We then construct two positive-probability histories of the same finite OPD process that reach one observation fiber but induce different recovery laws, establishing non-identifiability within a common algorithm and initialization. In contrast, equality of the full transition state guarantees the same recovery law under a common kernel. We further characterize smooth, bistable, critical, and collapse regimes in a scalar recursion and decompose finite-query intervention value into access to a recovery-relevant state and the signed reliability of its feedback. This theory leads directly to an empirical test: construct reachable states that match under a declared observation, continue them under a common policy, and intervene on hidden variables consumed by the future transition.

We implement this test at two complementary levels. **First**, an exact construction creates eight distinct Adam states at each of eight byte-identical DistilGPT2 weight states and evaluates 4,096 trajectories on common future tapes. Prespecified contrasts, an independent tape split, and permutation calibration all detect short-horizon recovery separation, while resetting the optimizer removes it. **Second**, we test whether optimizer state remains relevant under ordinary training with nonzero learning rates. Outcome-blind matching selects pairs with similar current functions, after which factorial cross-swaps separate optimizer-state and weight contributions in a procedural Transformer and in Qwen3.5 distillation on GSM8K. The exact construction provides the identifiability certificate, while these approximate matches test whether the same mechanism persists in practical training. Additional intervention and prospective-seed experiments separate feedback mechanisms, recovery-gap magnitude, and state ordering, while retaining uncertainty in effect signs and variation across query settings.

**Contributions.** We make three contributions:

- We define finite-budget recoverability, characterize observation sufficiency by fiber constancy, and derive a sharp minimax lower bound from residual fiber diameter.
- We construct positive-probability OPD histories with identical observations but different recovery laws, characterize scalar recovery regimes, and decompose finite-query intervention value into state access and signed teacher reliability.
- We realize an exact Transformer weight fiber with distinct Adam states and use optimizer cross-swaps under nonzero learning rates to test state-dependent recovery in procedural and Qwen3.5 GSM8K distillation.

## 2 Related Work

**On-Policy and Sequence Distillation.** For autoregressive models, the student determines the prefixes on which later supervision is requested. Scheduled sampling and Professor Forcing address the resulting train-test shift through generated prefixes or free-running dynamics (Bengio et al., 2015; Lamb et al., 2016), although changing the prefix source alone need not define a consistent sequence objective (Huszár, 2015). These approaches characterize distribution mismatch; our criterion asks whether equal output summaries determine future recoverability. GKD evaluates teacher feedback on student-generated sequences, MiniLLM derives an on-policy reverse-KL objective, and sequence-level divergence decompositions distinguish teacher and student prefix occupancies (Agarwal et al., 2024; Gu et al., 2024; Wen et al., 2023). Recent empirical work further emphasizes teacher-student compatibility and reliable token regions (Li et al., 2026; Xing et al., 2026). Our criterion connects these state-dependent effects to the information in current logits, entropy, divergence, and the full training state.

**Interactive Learning and Endogenous Sampling.** Complementing sequence distillation, interactive imitation learning provides an account of learner-induced data. DAgger queries an expert on states reached by the learner, and minimax analyses characterize when interaction improves statistical

rates (Ross et al., 2011; Rajaraman et al., 2021). Budget-aware querying and converging supervisors show that both access and teacher quality can vary during learning (Hoque et al., 2022; Balakrishna et al., 2020). Our analysis separates these roles: visitation controls whether a recovery-relevant state is observed, while conditional reliability controls the sign of the information obtained there. State-dependent performative prediction models learning as stochastic approximation driven by a learner-dependent controlled Markov chain (Li & Wai, 2022). Our finite recovery event and observation map connect this dynamical setting to recoverability identifiability when the map erases transition-relevant variables.

**State Abstraction and Training Memory.** More broadly, state abstraction formalizes what a representation preserves. Learned Markov abstractions and history-based information states identify conditions under which compressed observations retain future behavior (Allen et al., 2021; Subramanian et al., 2021; Majeed & Hutter, 2018). Bisimulation and its metric relaxations provide the corresponding transition-based language (Givan et al., 2003; Ferns et al., 2011). Our observation-fiber criterion is event-specific: it asks whether one recovery functional is constant within each fiber. OPD also contains algorithmic memory. Replay and target-network state already illustrate how future updates can depend on variables absent from current predictions (Schaul et al., 2015; Mnih et al., 2013). We connect these literatures by defining recoverability on the full training state, constructing two reachable OPD histories with the same observable summary, and separating natural visitation from direct state intervention.

## 3 Recoverability and State Identifiability

### 3.1 Problem Formulation

We model on-policy distillation as a controlled stochastic process. Its full state $x_t \in \mathsf{X}$ must determine both the next training example and the next update, including any persistent optimizer, buffer, teacher, sampler, or query-counter variables. A query and update policy $\mathcal{Q}$ specifies the distribution of teacher queries, returned information, and student updates, inducing the transition kernel $K_{\mathcal{Q}}(\mathrm{d}x_{t+1} \mid x_t)$.

One useful decomposition is

$$x_t = (\theta_t, m_t, b_t, c_t, u_t, \xi_t), \tag{1}$$

where $\theta_t$ contains student parameters, $m_t$ optimizer memory, $b_t$ replay or prompt-buffer state, $c_t$ teacher or calibration state, $u_t$ persistent update geometry, and $\xi_t$ random-tape and sampler state. These components organize variables by operational role; implementations can regroup them or omit variables the transition never reads. The defining requirement is closure: conditional on $x_t$ and $\mathcal{Q}$, the law of the next state and recovery flag is determined. Fig. 1(a) previews the construction used later: two reachable histories from one initialization can arrive at different full training states.

An observer sees a measurable summary $o_t = \phi(x_t)$, where $\phi : \mathsf{X} \to \mathsf{O}$ may retain probe logits, teacher divergence, entropy, support overlap, occupancy, or task success. We write $x \sim_\phi x'$ when $\phi(x) = \phi(x')$. This equality concerns the declared measurements and need not imply equality of the full state. Fig. 1(b) shows how such states can nevertheless share the same observable summary.

Fix a measurable recovery event and a finite budget $B$. The first time at which the event occurs is $\tau_{\mathrm{rec}}$. The budgeted recoverability functional is

$$\mathcal{R}_B(x; \mathcal{Q}) = \Pr_{x,\mathcal{Q}}(\tau_{\mathrm{rec}} \le B)\,. \tag{2}$$

Equation 2 includes future sampling, querying, teacher, and optimization randomness under $\mathcal{Q}$. Other bounded finite-horizon functionals, including normalized recovery area, admit the same analysis after augmenting the state with the accumulated statistic. Recoverability is always indexed by an event, policy, and budget.

For an observation $o$, define the observation fiber $\mathsf{F}_o = \{x : \phi(x) = o\}$. The projected recoverability diameter is

$$\Delta_B(o; \mathcal{Q}) = \sup_{x,x' \in \mathsf{F}_o} |\mathcal{R}_B(x; \mathcal{Q}) - \mathcal{R}_B(x'; \mathcal{Q})|\,. \tag{3}$$

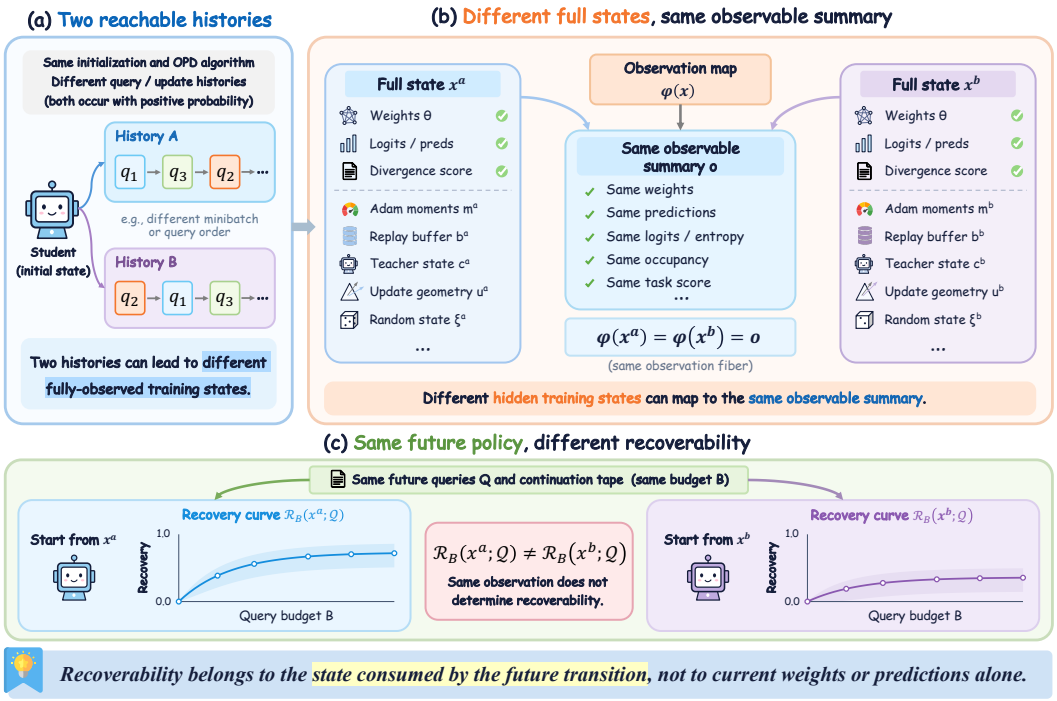


Figure 1. Full training states can share the same observable summary yet have different recovery laws. Panels (a)–(c) trace this ambiguity from reachable histories to future recovery.

The diameter measures ambiguity after observing $o$: zero means event-specific sufficiency on that fiber, while a positive value witnesses distinct recovery laws under the same summary. It therefore characterizes an information limitation independently of the estimator used. Fig. 1(c) shows why this ambiguity matters: a common future policy can produce different recovery laws from the same observation fiber.

Notation, randomized diagnostics, and measure-theoretic details are given in Appendix A.

Functional sufficiency means that $\mathcal{R}_B$ is constant on each observation fiber. Dynamical sufficiency requires the projection to determine the next projected state and recovery-flag law; it implies functional sufficiency at every finite horizon. Our empirical program tests the event-specific criterion using reachable histories, common continuations, and carrier interventions.

### 3.2 Full-State Sufficiency

We first establish the positive baseline: if the full transition state is identical, future recoverability is determined.

**Proposition 1 (full-state sufficiency).** If $x = x'$ as elements of the sufficient state space and both executions use the same kernel $K_{\mathcal{Q}}$, then $\mathcal{R}_B(x; \mathcal{Q}) = \mathcal{R}_B(x'; \mathcal{Q})$ for every finite budget $B$.

The coupling argument and implementation-level control are detailed in Appendix A.2.

### 3.3 Identifiability on Observation Fibers

Having established full-state sufficiency, we next ask when a coarser observable summary retains enough information to determine recoverability. For fixed $B$ and $\mathcal{Q}$, say that recoverability factors through $\phi$ if there is a measurable function $r_B : \mathsf{O} \to [0, 1]$ such that

$$\mathcal{R}_B(x; \mathcal{Q}) = r_B(\phi(x)) \quad \text{for every } x \in \mathsf{X}. \tag{4}$$

**Theorem 1 (observation-fiber characterization).** The factorization in Equation 4 exists if and only if $\Delta_B(o; \mathcal{Q}) = 0$ for every nonempty observation fiber. The diameter also controls the minimum

achievable error for estimates based only on $o$. Let $\widehat{r}(o)$ be any deterministic or randomized estimator whose input is only $o$. For any $x, x' \in \mathsf{F}_o$, the estimator has the same output law under both states. The triangle inequality gives

$$\begin{aligned}\mathbb{E}|\widehat{r}(o) - \mathcal{R}_B(x;\mathcal{Q})| + \mathbb{E}|\widehat{r}(o) - \mathcal{R}_B(x';\mathcal{Q})| \\ \geq |\mathcal{R}_B(x;\mathcal{Q}) - \mathcal{R}_B(x';\mathcal{Q})|\,. \end{aligned} \tag{5}$$

Taking the larger risk and then the supremum over the fiber yields the minimax lower bound

$$\inf_{\widehat{r}} \sup_{x \in \mathsf{F}_o} \mathbb{E}|\widehat{r}(o) - \mathcal{R}_B(x;\mathcal{Q})| \geq \frac{1}{2}\Delta_B(o;\mathcal{Q}). \tag{6}$$

### 3.4 Reachable-History Separation

The fiber criterion becomes substantive only if ambiguous states are reachable within the same learning process. We therefore construct two positive-probability histories from a common initialization.

**Theorem 2 (reachable-history non-identifiability).** There is a finite OPD process, a common initialization, a finite time $t$, and two positive probability histories $H_t^a, H_t^b$ such that $\phi(T(H_t^a)) = \phi(T(H_t^b))$. Under a common deterministic direct query continuation policy $\mathcal{Q}_{\mathrm{dir}}$ that forces the recovery relevant opportunity, one history has recovery probability zero and the other has recovery probability one. Every estimator restricted to the observable summary therefore has worst case absolute error at least one half. Under the natural on policy continuation $\mathcal{Q}_{\mathrm{nat}}$, the same two histories still have different recovery probabilities.

**Corollary 3.1** (Permutation fibers for adaptive optimizers)**.** *Fix a parameter vector $\theta_0$ and $N$ on policy minibatches whose gradients $g_1, \ldots, g_N$ are evaluated at $\theta_0$, with stochasticity attached to minibatch identity. Run Adam from a common optimizer state with zero construction learning rate. Every permutation $\pi$ ends at the exact weight state $\theta_0$, while its first moment satisfies*

$$m_N^{\pi} = (1-\beta_1)\sum_{t=1}^{N} \beta_1^{N-t} g_{\pi(t)} \tag{7}$$

*when the initial moment is zero. If two unequal gradients exchange positions with unequal exponential weights, their final first moments differ. The corresponding second moments can also differ. Thus permutations of one minibatch multiset can form an exact weight state fiber with distinct transition relevant optimizer states.*

The reachable construction and its natural-continuation separation are given in Appendix A.4.

### 3.5 Scalar On-Policy Dynamics

Non-identifiability describes what an observation can fail to determine; we next examine how on-policy visitation shapes recovery dynamics. Let $p = \sigma(\theta)$ be a branch probability. A length-$H$ opportunity occurs with probability $p^H$, where $H \geq 2$; its teacher target is $q \in (0,1]$, while a competing zero-target update has weight $\lambda \geq 0$. The stop-gradient mean field is

$$\dot{p} = p^2(1-p)\left[p^{H-1}(q-p) - \lambda\right], \qquad \lambda_{\max} = q^H \frac{(H-1)^{H-1}}{H^H}. \tag{8}$$

**Theorem 3 (scalar regime characterization).** For this field, $\lambda = 0$ gives positive drift on $(0, q)$; $0 < \lambda < \lambda_{\max}$ gives two interior roots, with attraction to zero below the lower root and to the upper root above it; $\lambda = \lambda_{\max}$ gives a critical double root at $p_\star = (H-1)q/H$; and $\lambda > \lambda_{\max}$ gives collapse toward zero. These are mean-field regimes; finite stochastic paths can cross the unstable root. Appendix A.5 derives the recursion and proves the characterization.

### 3.6 Natural Visitation and State Intervention

Finally, we separate whether a recovery-relevant state is reached from whether the feedback obtained there is useful. Suppose each natural query visits that state independently with probability $h$. Under a budget of $B$ queries, the probability of seeing it at least once is

$$V_B(h) = 1 - (1-h)^B. \tag{9}$$

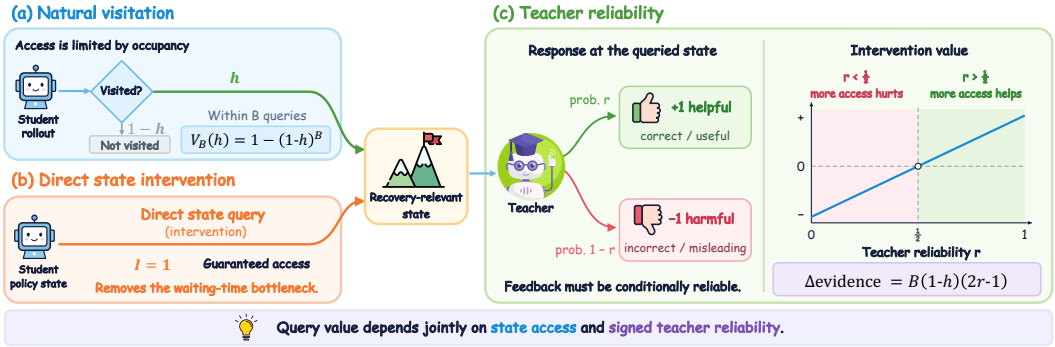


Figure 2. Natural visitation, direct intervention, and teacher reliability jointly determine query value. Panels (a)–(c) separate access from feedback sign.

Fig. 2(a) illustrates this occupancy-limited natural access.

Let $I_i \sim \text{Bernoulli}(h)$ indicate whether query $i$ reaches the recovery-relevant state. Direct state intervention sets $I_i = 1$ by explicitly presenting that state. Fig. 2(b) shows how direct intervention removes this waiting-time bottleneck.

Conditional feedback is represented by a signed response $Z_i \in \{-1, +1\}$ with $\Pr(Z_i = +1) = r$. Natural and direct access use the same conditional response law at that state, and $I_i$ is independent of $Z_i$. Fig. 2(c) isolates the remaining factor: teacher reliability determines whether additional access helps or harms. Natural visitation contributes $I_i Z_i$, while direct state intervention contributes $Z_i$. Under the same number of query slots, the difference in expected signed evidence is

$$\mathbb{E}\left[\sum_{i=1}^{B} Z_i - \sum_{i=1}^{B} I_i Z_i\right] = B(1-h)(2r-1). \tag{10}$$

**Theorem 4 (reliability-signed evidence value).** Direct state intervention contributes more expected signed evidence than natural visitation when $r > 1/2$, contributes the same amount when $r = 1/2$ or $h = 1$, and contributes less when $r < 1/2$. Under an update whose progress statistic is monotone in this signed sum, the same sign comparison holds for that statistic. A recovery-probability comparison requires an additional link between the progress statistic and the recovery event.

The theorem factors intervention value into access and response sign. Direct access increases the number of responses at the target state, while teacher reliability determines their expected sign. Query-complexity, concentration, and scalar-dynamics details appear in Appendices A.5 and A.6.

## 4 Experimental Design

We mirror the theoretical pipeline with three complementary evaluations: an exact identifiability certificate, nonzero-learning-rate carrier tests, and prospective and interventional validation.

### 4.1 Exact Weight Fiber with Optimizer Memory

The exact-weight experiment isolates optimizer memory in a full-parameter causal Transformer. For each of eight seeds, a two-update warmup produces one DistilGPT2 base weight state and a fresh procedural task with 16 states, two student-sampled branches, and programmatic teacher targets. From that weight state, we generate 12 on-policy minibatches once and process the same multiset in eight orders: identity, reverse, and six fixed hash-derived permutations. AdamW uses zero learning rate during this construction, so gradients update its first and second moments while student parameters and buffers remain unchanged. Dropout randomness is keyed by minibatch identity, making each minibatch gradient independent of its position. We retain complete weight, optimizer, order, and minibatch-multiset hashes.

Each hidden optimizer state is evaluated on the same 64 future random tapes. A tape fixes procedural prompts, sampling uniforms, and dropout randomness, while generation remains on-policy because

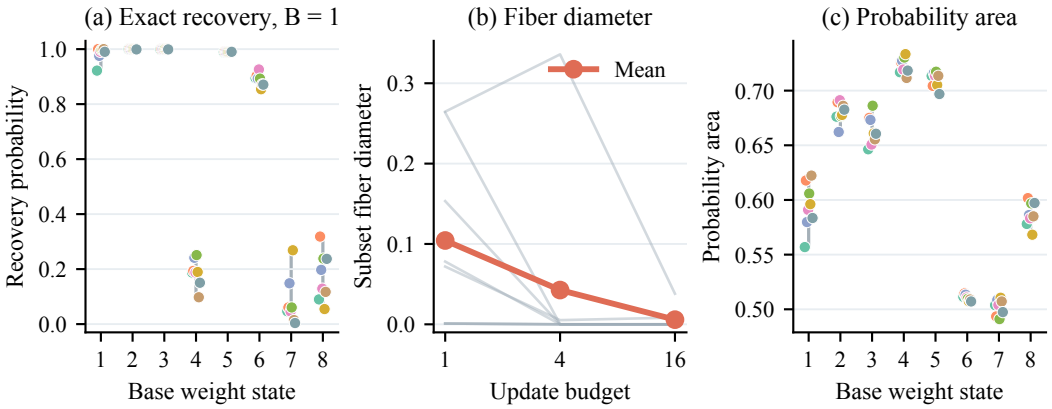


Figure 3. Recovery across eight hidden Adam states at byte-identical weight states under 64 common future tapes. (a) One-step exact recovery. (b) Subset fiber diameter across budgets. (c) Normalized recovery area.

each student maps the common uniforms through its own distribution. Recovery is evaluated at budgets 1, 4, and 16 with learning rate $5 \times 10^{-5}$, matching the warmup rate. The recovery event is exact top-1 agreement with the programmatic teacher over two admissible answer tokens on every state initially answered incorrectly. We also report normalized area under the teacher-target probability curve. Pilot-rate and estimator details are given in Appendix B.2.

### 4.2 Nonzero Learning Rate Natural Histories

We next test whether optimizer state remains relevant under ordinary training with nonzero learning rates. We use the same causal design in a procedural Transformer and in Qwen3.5 distillation on GSM8K. Candidate histories are generated from a common base state by ordinary on-policy updates, and pair selection uses only the current student function before any recovery outcome is observed. Each selected pair yields the natural states $(\theta_A, m_A)$ and $(\theta_B, m_B)$, the cross-swaps $(\theta_A, m_B)$ and $(\theta_B, m_A)$, and a common zero-moment control. The four crossed states give the tape-level decomposition

$$D_{\text{natural}} = D_{\text{optimizer}} + D_{\text{weights}}. \tag{11}$$

The exact-weight experiment supplies the identifiability certificate; these approximate function matches test whether the same carrier remains active during ordinary training.

### 4.3 Prospective and Interventional Validation

We separately test whether recovery-gap magnitude and state ordering transfer to unseen seeds, using frozen predictions from the main census. We also evaluate interventions that alter access to recovery-relevant states or the feedback delivered there while preserving common query and update accounting. These experiments distinguish group-level recovery magnitude, state-level ordering, and the signed value of additional access. Full protocols, estimators, outcome definitions, and extended results are provided in Appendix B.

## 5 Results and Analysis

The results follow the same sequence: we first establish exact non-identifiability and identify its optimizer carrier, then test that carrier under ordinary training, and finally examine prospective transfer and intervention value.

**Observation 1: Identical current weights can lead to different recovery probabilities.** We begin with the exact weight fiber. Within each base seed, all eight histories have byte-identical parameters, so every deterministic current-model query agrees. Nevertheless, their recovery probabilities differ under a common continuation (Fig. 3). At budget one, the prespecified identity-minus-reverse contrast is $-0.042$ with base-seed interval $[-0.099, -0.004]$. Histories selected on the first 32 tapes remain

Table 1: Procedural setting of the nonzero learning rate natural history factorial. Each entry is the mean tape sign flip calibrated absolute contrast with a base seed bootstrap interval. Natural is the matched history contrast. Optimizer and weights are the two factorial main effects from cross swapping Adam states. Reset uses a common zero moment state.

| Functional | Natural | Optimizer | Weights | Reset |
|---|---|---|---|---|
| Exact, $B = 1$ | $.044\,[.001, .109]$ | $.057\,[.005, .134]$ | $.010\,[.000, .021]$ | $.038\,[-.002, .097]$ |
| Exact, $B = 4$ | $.132\,[.090, .171]$ | $.146\,[.098, .187]$ | $.035\,[.005, .071]$ | $.025\,[.011, .039]$ |
| Exact, $B = 16$ | $.023\,[.001, .050]$ | $.009\,[-.001, .023]$ | $.014\,[.002, .026]$ | $.001\,[-.004, .005]$ |
| Probability area | $.016\,[.010, .023]$ | $.009\,[.005, .013]$ | $.010\,[.003, .017]$ | $.006\,[.003, .009]$ |

Table 2: Qwen3.5 GSM8K setting of the nonzero learning rate natural history factorial. Each entry is a base seed mean with a bootstrap interval. Probability rows ($p_{\mathrm{ref}}$) are in units of $10^{-3}$; the exact-recovery row is unscaled. Signed effects use the outcome blind first moment norm orientation. The last two columns are absolute effects calibrated by a tape sign flip null.

| Functional | Adam signed | Weights signed | Adam abs. | Weights abs. |
|---|---|---|---|---|
| Exact, $B = 1$ | $.138\,[-.033, .301]$ | $-.004\,[-.089, .096]$ | $.163\,[.069, .259]$ | $.052\,[.009, .103]$ |
| $p_{\mathrm{ref}}$, $B = 4$ | $6.9\,[0.1, 15.2]$ | $-0.4\,[-3.2, 2.8]$ | $6.9\,[2.3, 13]$ | $1.4\,[-0.2, 3.2]$ |
| $p_{\mathrm{ref}}$, $B = 8$ | $5.9\,[0.7, 11.4]$ | $-0.1\,[-3, 3.1]$ | $4.4\,[0.8, 8.7]$ | $1.5\,[0.2, 3]$ |
| $p_{\mathrm{ref}}$ area | $4.4\,[-0.2, 10]$ | $0.2\,[-2.7, 3.1]$ | $4.1\,[0.7, 8.6]$ | $2.1\,[0.9, 3.4]$ |

separated on the other 32, yielding a split estimate of 0.100 [0.033,0.175]. The raw diameter is 0.104, the permutation-null mean is 0.016, and the calibrated diameter is 0.088. Continuous recovery area likewise has split separation 0.026 [0.015,0.036] and calibrated diameter 0.022 [0.012,0.033]. Separation contracts at larger budgets as cumulative recovery approaches its ceiling. Thus, the certificate is specific to a declared event, policy, and budget; Appendix B.2 and Table 4 give the complete estimators and results.

**Observation 2: Equalizing optimizer memory removes the exact-weight separation.** We next test the carrier responsible for this ambiguity. Zeroing both Adam moments maps the eight histories to one optimizer state because parameter values, parameter order, and step counts already agree. Their common-tape recoveries then agree exactly. Stateless SGD provides a complementary control: under zero-learning-rate construction, it retains no order-dependent optimizer memory, and its paired recoveries also agree exactly. These controls keep examples, teacher targets, weights, and future randomness fixed while removing inherited optimizer memory. Equalizing the continuation state therefore restores the same recovery law, as required by Proposition 1.

**Observation 3: Optimizer effects persist near function matches after ordinary training.** We then move from the exact certificate to histories generated with nonzero learning rates. Eight procedural pairs are selected from 1,024 histories using only their current student functions. Their mean maximum log-probability distance is 0.133, RMS distance is 0.048, and relative parameter distance is $6.88 \times 10^{-4}$; the selected function distance is 15.5% of the candidate-pair median. Because these are approximate rather than exact matches, factorial cross-swaps separate optimizer effects from residual weight differences. At budget one, the tape-sign-flip-calibrated optimizer magnitude is 0.0570 [0.0055,0.1345], compared with a weights magnitude of 0.0098 and natural paired magnitude of 0.0437 (Table 1). The optimizer magnitude reaches 0.1461 [0.0981,0.1872] at budget four and remains 0.0087 [0.0051,0.0126] for continuous recovery area. Optimizer relevance therefore extends beyond the zero-learning-rate certificate, while exact fiber equality and approximate function matching remain distinct claims.

**Observation 4: Qwen3.5 histories show optimizer sensitivity with uncertain signed effects.** The same design on GSM8K yields seven valid base seeds from eight declared seeds and 1,344 recovery trajectories. The selected current-function $L_\infty$ distance averages 1.089, or 0.394 times the within-seed candidate-pair median, with RMS distance 0.311. At budget one, the exact-recovery optimizer effect has signed mean 0.138 [$-0.033$,0.301] and calibrated absolute magnitude 0.163 [0.069,0.259]; the corresponding weight effects are $-0.004$ [$-0.089$,0.096] and 0.052 [0.009,0.103] (Table 2). At budgets four and eight, signed optimizer effects on reference-answer probability are

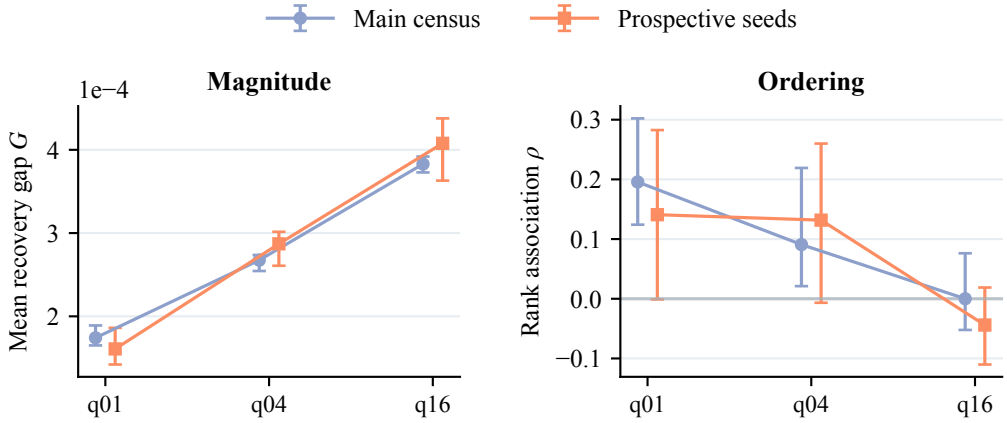


Figure 4. Prospective transfer of recovery-gap magnitude $G$ and rank association $\rho$. Points show group means and bars show seed ranges; main-census means are frozen predictions.

0.0069 and 0.0059, whereas weight effects are $-0.0004$ and $-0.0001$. Exact base-seed sign-flip probabilities for the three signed optimizer contrasts are 0.219, 0.141, and 0.125. The results therefore support state-dependent optimizer sensitivity without establishing a uniform signed advantage across seeds. Appendix B.3 gives the retained events and controls.

**Observation 5: Recovery-gap magnitude transfers more consistently than state ordering.** We next compare frozen ten-seed census predictions with prospective seeds (Fig. 4; Table 5). New-seed recovery gaps remain on the original scale: q04 and q16 means are slightly above their frozen predictions, whereas q01 is slightly below. By contrast, the q16 mean rank association changes sign despite similar gap magnitude. Recovery variation among matched states therefore establishes ambiguity, while ordering individual histories is a separate prediction problem. Omitted carriers can act in opposing directions and change influence across query settings, so magnitude transfer need not imply ranking transfer. Appendix C.1 reports the full ranges and prospective diagnostics.

**Observation 6: More access helps only when the feedback it exposes is useful.** Finally, the intervention atlas separates access from response sign (Fig. 5; Table 6). Directional targets and distributional guidance are positive across seeds, while rollout exposure and the hazard variant are negative; ancestor queries have mixed signs. Continued sampling changes from negative at q01 to positive at larger query budgets. These signed contrasts are preserved under common query and update accounting and are consistent with Theorem 4: increasing access amplifies the conditional value of the feedback obtained there. Top-$k$ and full-distribution guidance coincide because the top-$k$ payload is saturated and therefore constitute one effective information contrast. The atlas captures collective hidden-state effects, while the exact-weight reset in Observation 2 provides the specific optimizer attribution. Appendix C gives additional controls and neural-panel coverage.

## 6 Conclusion

We develop a state-based account of recoverability in on-policy distillation: future recovery is determined by the state consumed by subsequent sampling, querying, and updating, not by current weights or predictions alone. Our observation-fiber criterion characterizes exactly when a reduced summary preserves recoverability and quantifies the ambiguity left when it does not. Exact Transformer constructions demonstrate that byte-identical weights with different Adam states can induce different recovery laws, while optimizer resets remove this separation. Nonzero-learning-rate experiments further establish optimizer-state relevance under ordinary training, with signed effects varying across histories and settings. Prospective and interventional results additionally separate recovery-gap magnitude from state ordering, and state access from feedback reliability. This perspective shifts recovery analysis from comparing current models to testing whether the observed training state is sufficient for the future transition. Together, these results provide a framework for testing which aspects of training state are necessary to predict future learning under a declared event, policy, and budget.

## AI USAGE STATEMENT

Generative AI systems (GPT-6 Astra and Codex) are used as limited editorial and implementation aids, including code drafting and debugging, language polishing, and manuscript organization. The authors design the mathematical framework and experiments, evaluate the scientific interpretations, and make the final decisions regarding every claim. The systems used to prepare the manuscript do not generate the experimental evidence or replace the analyses reported here. We inspect and test AI-assisted code against the intended computations, check numerical statements against the recorded outputs, and review and revise all AI-assisted prose. The authors therefore retain responsibility for the final text, mathematics, results, software, and accompanying materials.

## REFERENCES

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

Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8229–8241, 2021.

Ashwin Balakrishna, Brijen Thananjeyan, Jonathan Lee, Felix Li, Arsh Zahed, Joseph E. Gonzalez, and Ken Goldberg. On-policy robot imitation learning from a converging supervisor. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 24–41. PMLR, 2020.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*, 2015.

Gaoang Bian and Alessandro Abate. On the relationship between bisimulation and trace equivalence in an approximate probabilistic context. *arXiv preprint arXiv:1701.04547*, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Wojciech M. Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1331–1340. PMLR, 2019.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023. doi: 10.1287/moor.2022.1287.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 598–608. PMLR, 2022.

Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *arXiv preprint arXiv:1610.09038*, 2016.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3164–3186. PMLR, 2022.

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

Sultan Javed Majeed and Marcus Hutter. Performance guarantees for homomorphisms beyond markov decision processes. *arXiv preprint arXiv:1811.03895*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nived Rajaraman, Yanjun Han, Lin Yang, Jingbo Liu, Jiantao Jiao, and Kannan Ramchandran. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1325–1336, 2021.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *arXiv preprint arXiv:2010.08843*, 2021.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. $f$-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

# A Separation and Decision Theory

## A.1 Notation

Table 3 collects the objects used in the analysis. The distinction between $x$ and $o$ is maintained throughout. The state $x$ is sufficient for the declared transition kernel by construction, while $o$ is a candidate summary whose sufficiency must be established rather than assumed.

The observation map can be deterministic or randomized. A randomized diagnostic can be incorporated by augmenting $x$ with its probe tape and treating the realized report as $o$. This prevents repeated probing noise from being mistaken for hidden training state. The map can also depend on a fixed probe distribution, but that distribution must be part of the declaration. Logit equality on one prompt set and logit equality on all reachable prefixes define different fibers. The recovery diameter is meaningful only relative to the chosen fiber.

We take $\mathsf{O} = \phi(\mathsf{X})$ and equip this image with the final sigma algebra

$$\mathcal{O}_\phi = \{A \subseteq \mathsf{O} : \phi^{-1}(A) \text{ is measurable in } \mathsf{X}\}. \tag{12}$$

This convention makes the measurable structure of an observation explicit and avoids imposing an unproved regularity property on an arbitrary ambient report space.

## A.2 Full-State Sufficiency

The first statement is a coupling sanity check. Suppose two executions begin from the same full state $x$ and use the same policy $\mathcal{Q}$. Couple their query choices, teacher responses, minibatches, update noise, and random tapes identically. Because both executions apply the same transition kernel to the same state, induction gives equal states at every future step until the budget ends. Their recovery indicators are equal pathwise, hence their recovery probabilities agree.

The proposition concerns equality in the declared sufficient state space. This equality includes every variable read by the future transition. Optimizer momentum, replay cursors, teacher state, and random generators enter whenever they affect the next update, while components outside the future kernel can be omitted. The correct state is therefore determined by operational dependence.

Proposition 1 also distinguishes state equality from exchangeability. Two states sampled from the same distribution can have equal marginal recovery probabilities without being pathwise equal under a common tape. Conversely, two distinct byte representations can define the same transition law after a change of coordinates. For the theory, the relevant equality is equality as sufficient states of the declared kernel. For the implementation control, canonical serialization provides a conservative realizable test: it enforces identical values for all named carriers and rejects unrecorded state. A representation-level equivalence could be weaker, but it would require its own proof that the kernel and recovery event are invariant under the representation map.

The proof also explains the full-state sufficiency control used in the experiments. Two histories are generated through independent execution paths, after which all declared state carriers are canonicalized to identical values. They are then evolved under a common query and random tape. Any residual difference would expose a missing carrier, nondeterministic transition, or implementation defect before it could be interpreted scientifically. Exact equality in this control is therefore an implementation-level realization of Proposition 1 for the named finite process.

Table 3: Notation used in the state-based recoverability analysis.

| Symbol | Type | Meaning |
|---|---|---|
| $x \in \mathsf{X}$ | State | Full training state of the declared OPD process |
| $o = \phi(x)$ | Summary | Observable training summary |
| $\mathsf{F}_o$ | Set | Full states mapped to observation $o$ |
| $\mathcal{Q}$ | Policy | Teacher-query and student-update policy |
| $B$ | Integer | Finite query and update budget |
| $\eta_t$ | Scalar | Nonnegative student-update step size |
| $\mathcal{R}_B$ | Probability | Budgeted recoverability |
| $\Delta_B$ | Scalar | Projected recoverability diameter |
| $p$ | Probability | Natural visitation or scalar branch probability |
| $r$ | Probability | Teacher reliability on the queried state distribution |
| $G$ | Statistic | Mean absolute paired recovery-curve difference |
| $\rho$ | Statistic | Tie-aware rank association with the recovery gap |

This coupling perspective follows established ideas from stochastic bisimulation and Markov state abstraction (Givan et al., 2003; Ferns et al., 2011). Here the equivalence is indexed by the declared recovery event and policy: equalized full states induce the same future kernel and recovery event under that policy. This condition is weaker than policy-uniform reward and transition equivalence.

### A.3 Identifiability on Observation Fibers

The forward direction is immediate because two states in one fiber have the same argument to $r_B$. For the reverse direction, zero diameter makes $\mathcal{R}_B$ constant on each fiber, so define $r_B(o)$ to be that constant. For every Borel set $C \subseteq [0,1]$, we have $\phi^{-1}(r_B^{-1}(C)) = \mathcal{R}_B^{-1}(C)$. Measurability of $\mathcal{R}_B$ and Equation 12 therefore imply measurability of $r_B$. The theorem characterizes universal correctness for predictors restricted to $o$; constructing or learning $r_B$ is the corresponding statistical problem.

When the lower and upper envelopes of recoverability on the fiber are attained, their midpoint reaches the right side of Equation 6. The bound is therefore sharp under absolute loss. Its meaning is operational: ambiguity inside the fiber becomes unavoidable estimation risk, even with unlimited data about the observation itself. More data can estimate the mixture induced by a population over hidden states, but it cannot identify which recovery law applies to an individual state unless the data reveal information beyond $o$.

This point separates population calibration from state identification. Let $\mu(\mathrm{d}x \mid o)$ be a population distribution over hidden states in the fiber. The conditional mean $\bar{r}(o) = \int \mathcal{R}_B(x; \mathcal{Q})\mu(\mathrm{d}x \mid o)$ can be estimated accurately from repeated data drawn from that population. It is optimal for squared prediction error under $\mu$. Yet the same estimator can have large statewise error when the population mixes distinct recovery laws. The fiber diameter is distribution free, while population risk depends on the mixing weights. Our empirical $G$ statistic targets paired within-fiber variation, and the prospective prediction error targets population stability across new seeds. They answer complementary questions.

The criterion also supports representation comparison. Suppose $\phi_1$ refines $\phi_0$, so every $\phi_1$ fiber lies inside a $\phi_0$ fiber. Then projected recoverability diameter cannot increase under refinement:

$$\Delta_B^{\phi_1}(\phi_1(x); \mathcal{Q}) \le \Delta_B^{\phi_0}(\phi_0(x); \mathcal{Q}) \quad \text{whenever the comparison is made at } x. \tag{13}$$

Equation 13 orders summaries by information. Richer summaries reduce the set of hidden recovery values compatible with a specific state while increasing storage, privacy exposure, and estimation complexity. A practical summary-selection problem can therefore trade within-fiber ambiguity against measurement cost.

### A.4 REACHABLE-HISTORY SEPARATION

Reachability is part of the certificate. Both states must arise with positive probability from the same initialization and the same finite update algorithm. Let $H_t$ denote the realized history up to time $t$, including sampled training events. A history map $T$ sends $H_t$ to the full training state $x_t = T(H_t)$. Two histories can produce equal current parameters while their optimizer memories differ because different gradient sequences can cancel in the parameter coordinate without canceling in momentum or second-moment coordinates.

Here is an exact construction. Measure a scalar logit in units of $\log 3$, write the full state as $x_k = (\theta_k, m_k)$, and set $p_k = \sigma(\theta_k \log 3)$. Start from $(\theta_0, m_0) = (0, 0)$. A length-two student trajectory exposes a teacher opportunity $Y_k \in \{0, 1\}$ with conditional probability $p_k^2$. Given teacher target $q = 3/5$, momentum factor $\beta = 1/2$, and step size $\eta = 10$, one OPD update is

$$g_k = Y_k(p_k - q), \qquad m_{k+1} = \frac{1}{2} m_k + g_k, \qquad \theta_{k+1} = \theta_k - 10 m_{k+1}. \tag{14}$$

The two realized histories are $H_2^a = (Y_0, Y_1) = (1, 1)$ and $H_2^b = (0, 0)$. Along $H_2^a$, the successive states are $(0, 0)$, $(1, -1/10)$, and $(0, 1/10)$. Along $H_2^b$, both updates leave the state at $(0, 0)$. Their probabilities under the same sampling law are

$$\Pr(H_2^a) = \left(\frac{1}{2}\right)^2 \left(\frac{3}{4}\right)^2 = \frac{9}{64} > 0, \qquad \Pr(H_2^b) = \left(1 - \left(\frac{1}{2}\right)^2\right)^2 = \frac{9}{16} > 0. \tag{15}$$

Let the observation be $\phi(\theta, m) = (\theta, p, k, q)$, so both time two states have observation $(0, 1/2, 2, 3/5)$ while their momentum coordinates differ. The direct continuation $\mathcal{Q}_{\mathrm{dir}}$ explicitly presents the recovery relevant teacher opportunity, which sets $Y_2 = 1$ for both states instead of sampling it from natural occupancy. Equation 14 gives $\theta_3^a = 1/2$ and $\theta_3^b = 1$. For the recovery event $p_3 \geq 2/3$ and budget one, $\sigma((\log 3)/2) < 2/3$ whereas $\sigma(\log 3) = 3/4$. Hence $\mathcal{R}_1(x_2^a; \mathcal{Q}_{\mathrm{dir}}) = 0$ and $\mathcal{R}_1(x_2^b; \mathcal{Q}_{\mathrm{dir}}) = 1$. Their direct policy fiber diameter is one, and Equation 6 gives worst case absolute error at least one half.

The natural continuation $\mathcal{Q}_{\mathrm{nat}}$ instead preserves $\Pr(Y_2 = 1 \mid p_2) = p_2^2 = 1/4$. History $a$ never reaches the event in one update, while history $b$ reaches it exactly when $Y_2 = 1$. Therefore

$$\mathcal{R}_1(x_2^a; \mathcal{Q}_{\mathrm{nat}}) = 0, \qquad \mathcal{R}_1(x_2^b; \mathcal{Q}_{\mathrm{nat}}) = \frac{1}{4}, \qquad \Delta_1(o; \mathcal{Q}_{\mathrm{nat}}) = \frac{1}{4}. \tag{16}$$

Equation 6 then gives a natural policy lower bound of $1/8$. The separation therefore persists under natural occupancy, while direct intervention changes its magnitude from one quarter to one.

The construction has three defining properties. First, both histories use a single initialization and algorithm. Second, observable equality holds before the future intervention. Third, the hidden variable is an ordinary training state component with a clear transition role. Its conclusion is event and policy specific: an optimizer difference is recovery relevant when it changes the declared recovery law, and any observation map that omits this component requires a sufficiency argument for that event and policy.

The parameter statement follows because every Adam parameter increment is multiplied by the zero construction learning rate. For positions $t \neq s$, exchanging $g_i$ and $g_j$ changes the first moment by

$$(1 - \beta_1)\left(\beta_1^{N-t} - \beta_1^{N-s}\right)(g_i - g_j), \tag{17}$$

which is nonzero whenever the displayed factors are nonzero. Each ordering has positive probability under repeated sampling whenever its constituent minibatches do. The exact weight experiment instantiates this corollary with ordinary Transformer gradients and then tests whether the resulting optimizer states separate a declared recovery law.

The same reasoning extends beyond momentum. A replay buffer can store examples that determine future gradients. A teacher state can change the distribution or sign of feedback conditional on the same student output. A local geometry cache can change a preconditioned update. In each case the criterion is counterfactual: hold the observable summary fixed, vary one omitted carrier through

reachable histories or a controlled intervention, and compare $\mathcal{R}_B$. A component is recovery relevant only if this operation changes the recovery law.

Carrier interventions require a consistency condition. Replacing one component of Equation 1 must produce a state in the domain of the same future kernel. We enforce this condition in two ways. The main impossibility result uses naturally reached states, and empirical carrier tests copy values between source-valid states while holding the remaining named fields and future tape fixed. A nonzero effect identifies carrier participation in the recovery law for the tested pair; a zero effect records invariance of the declared event under that intervention, policy, and budget.

The interventions also induce a directed dependence graph. A carrier can affect query selection, teacher response, update direction, or recovery scoring. Effects mediated solely through the current observable are blocked by matching $o$. Effects that remain must travel through an omitted path. This graph suggests a diagnostic sequence: begin with the full-state control, then restore candidate carriers one at a time, then test combinations only when individual effects do not explain the residual. The order avoids attributing a numerical mismatch to a hidden variable before establishing that equal declared state actually removes it.

Reachability also separates the theorem from generic representation-level history dependence. History-based information states can preserve future laws by retaining a suitable statistic (Subramanian et al., 2021; Majeed & Hutter, 2018). Our contribution is to identify the corresponding statistic for a concrete OPD transition and to show that a commonly inspected policy summary can omit it. The theorem is a possibility result for finite processes. Whether a named neural protocol exhibits a detectable diameter is an empirical question addressed separately.

### A.5 Scalar On-Policy Dynamics

The scalar model makes opportunity endogenous.

The separation theorem concerns information. A second question is dynamical: when can student-induced visitation create qualitatively different recovery regimes? Consider a scalar student probability $p_t = \sigma(\theta_t)$ for a recovery-relevant branch, and let $H \geq 2$ be an integer. A successful teacher opportunity requires a length-$H$ student path and occurs with probability $p_t^H$. Conditional on an opportunity, the teacher target is $q \in (0, 1]$. A competing update with nonnegative weight $\lambda$ favors the zero target. We differentiate the local cross-entropy losses while treating the sampled opportunity as a stop-gradient event, which matches the semi-gradient structure analyzed in policy distillation (Czarnecki et al., 2019).

Let $Y_t \sim \mathrm{Bernoulli}(p_t^H)$ be the opportunity indicator and let $\eta_t$ be the nonnegative step size. The scalar stochastic recursion is

$$\theta_{t+1} = \theta_t + \eta_t \left[ Y_t(q - p_t) - \lambda p_t \right]. \tag{18}$$

The two derivative terms are not postulated drifts. They are the exact derivatives of binary cross entropy with teacher targets $q$ and zero. Taking the conditional expectation of Equation 18 gives

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid p_t = p] = \eta_t \left[ p^H(q - p) - \lambda p \right]. \tag{19}$$

For a small-step mean-field analysis, multiplication by the sigmoid derivative gives

$$\dot{p} = p^2(1 - p) \left[ p^{H-1}(q - p) - \lambda \right]. \tag{20}$$

Define $\psi(p) = p^{H-1}(q - p)$. Its unique interior maximizer and maximum are

$$p_\star = \frac{H - 1}{H} q, \qquad \lambda_{\max} = q^H \frac{(H - 1)^{H-1}}{H^H}. \tag{21}$$

The proof follows from the unimodality of $\psi$. Its derivative has the sign of $(H - 1)q - Hp$, so a horizontal line at $\lambda$ intersects $\psi$ twice below its maximum, once at the maximum, and not at all above

it. The remaining factor $p^2(1-p)$ is nonnegative. The sign of the bracket therefore determines the phase portrait. The reference certificate sets $H = 2$, $q = 3/4$, and $\lambda = 1/8$. It obtains $p_\star = 3/8$, $\lambda_{\max} = 9/64$, and roots at $1/4$ and $1/2$, with negative, positive, and negative drift across the resulting intervals.

This theorem distinguishes three mechanisms that can look similar in a short run. In the smooth regime, small $p$ creates a long waiting time because the opportunity probability scales as $p^H$, but recovery drift remains positive. In the bistable regime, the lower root is a genuine barrier for the mean field. In the collapse regime, the conflict term dominates every interior state. Continuous recovery curves and signed intervention effects distinguish rare observation from negative expected drift and connect finite trajectories to this phase portrait.

The mean field describes expected local motion. A finite stochastic trajectory can cross an unstable root through noise, and a decreasing step schedule can prevent such crossings after the updates become small. The basin statement is exact for the deterministic field and guides the stochastic approximation analysis. The observable probability $p$ supplies one state coordinate; momentum or adaptive preconditioning can make the discrete recursion higher order even when the emitted Bernoulli probability is scalar.

The parameters have distinct interpretations. Increasing $H$ sharpens the occupancy bottleneck because an opportunity requires a longer successful prefix. Increasing $q$ strengthens the teacher pull conditional on an opportunity. Increasing $\lambda$ strengthens a competing update that is applied regardless of whether the recovery state is visited. The ratio between the peak of $\psi$ and $\lambda$ governs the regime. This resembles sensitivity-to-curvature conditions in decision-dependent optimization (Drusvyatskiy & Xiao, 2023), but the present recursion is derived from a particular stop-gradient OPD loss and has a directly interpretable visitation exponent.

The scalar model isolates the interaction between on-policy sampling and a conflicting update. Its empirical role is to generate qualitative questions about gap growth with query depth, intervention signs, and recovery differences after observable matching and full-state controls.

### A.6 Natural Visitation and State Intervention

To reach visibility at least $1 - \delta$, the natural-query budget must satisfy

$$B \geq \frac{\log \delta}{\log(1-h)}. \tag{22}$$

Equation 22 characterizes access. This distinction is consistent with trust-region and reliability-gated OPD methods, which restrict supervision when teacher and student distributions disagree (Xing et al., 2026).

Natural and interventional query complexity can now be stated in a common language. For a target visibility $1-\delta$, define $C_{\mathrm{nat}}(h, \delta)$ as the smallest natural-query budget satisfying Equation 22. If direct access is available, the corresponding access complexity is one query. This comparison concerns access only. To obtain correct signed evidence with probability $1 - \delta$, both complexities also depend on $r$ through concentration. When $r$ is close to one half, direct access removes the occupancy bottleneck but cannot remove label uncertainty. When $r$ is below one half, more direct queries concentrate the wrong sign.

The independence assumptions make the finite expressions transparent. Correlated visitation can be handled by replacing Equation 9 with a hitting-time bound for the actual state process. State-dependent reliability can be handled by conditioning $r$ on the visited or intervened state and summing predictable signed increments. The central distinction remains: visitation controls whether information is obtained, and conditional reliability controls what that information does. Conflating the two can make an intervention appear ineffective when the actual problem is rare access, or appear useful when it merely repeats biased supervision.

A finite concentration statement follows without asymptotics. Conditional on $N$ visible responses, Hoeffding's inequality bounds the probability that their signed sum has the wrong sign. Averaging over $N \sim \mathrm{Binomial}(B, h)$ yields

$$\Pr(\text{correct signed evidence}) \geq 1 - \exp\left[-Bh\left(1 - \exp\left(-\frac{(2r-1)^2}{2}\right)\right)\right]. \tag{23}$$

The exact finite certificate uses $B = 16$, $h = 1/4$, and $r = 3/4$. It verifies an expected intervention difference of 6 and shows that visibility $1 - \delta$ with $\delta = 0.05$ requires 11 natural queries when $h = 1/4$. These values instantiate Equations 22 and 10, whose sign law links intervention value to teacher reliability.

### A.7 Approximate Dynamical Sufficiency

Exact equality is useful for controls, but learned summaries are usually approximate. Augment the state with a binary recovery flag. Assume uniformly that, for every relevant pair of states with the same observation and recovery flag, their one-step kernels admit a history-adapted coupling under which the next flags agree and the next states remain observationally equivalent with conditional probability at least $1-\varepsilon$. Applying this coupling recursively whenever equivalence has been preserved gives a joint path that remains equivalent for $H$ steps with probability at least $(1-\varepsilon)^H$. Consequently,

$$\Delta_H(o; \mathcal{Q}) \leq 1 - (1-\varepsilon)^H \leq \min\{1, H\varepsilon\}. \tag{24}$$

If the initial states enter a common fiber only with probability at least $1 - \delta$, the corresponding bound is

$$\Delta_H(o; \mathcal{Q}) \leq 1 - (1-\delta)(1-\varepsilon)^H \leq \min\{1, \delta + H\varepsilon\}. \tag{25}$$

The nonlinear law in Equations 24 and 25 specializes finite-trace coupling bounds from approximate probabilistic bisimulation to recovery events (Bian & Abate, 2017). It shows that empirical matching must control both current observation error and the probability that coupled future states remain equivalent under the actual query and update kernel.

This observation explains why pointwise closeness of logits can leave dynamical ambiguity. A small current discrepancy can be followed by different minibatch distributions, optimizer responses, or teacher outputs, while a coarse observation can be sufficient for one short-horizon event. The empirical target is the recoverability variation left by the measured projection under the declared policy and budget.

The approximate result provides a design rule for diagnostics. A matching procedure should prioritize quantities that reduce the coupled next-step failure probability. An optimizer moment that changes the next update contributes directly to $\varepsilon$, while a probe statistic that determines the query rule can reduce coupling failure through the observation map. Our stress tests therefore pair observable matching with interventions and full-state controls to estimate different parts of the same dependence structure.

The finite-horizon form also discourages a purely local interpretation. A one-step coupling error can accumulate over repeated updates, while an event that is decided early can be insensitive to later divergence. Reporting $H\varepsilon$ without the nonlinear term can be loose when the horizon is long, and reporting only current observation error ignores transition amplification. For this reason, recovery curves are evaluated over the declared budget and summarized by area rather than by a single immediate derivative.

Together, exact and approximate sufficiency give two positive certificates: exact transition closure, as in the full-state control, and a reduced observation with controlled finite-horizon coupling error. Positive fiber diameter has its own certificate through reachable states in one fiber with separated recovery values. These certificates distinguish exact sufficiency, approximate sufficiency, and observable-state nonidentifiability.

### A.8 Empirical Functionals

The experiments evaluate a continuous finite-horizon functional. At steps $t = (0, 1, 2, 4, 8, 16)$, a fixed probe set gives the bounded score $c_t = \exp(-\mathrm{CE}_t)$. For one history $h$, define the normalized recovery area

$$A(h) = \sum_{j=0}^{4} \frac{t_{j+1} - t_j}{t_5 - t_0} \frac{c_{t_j}(h) + c_{t_{j+1}}(h)}{2}. \tag{26}$$

For matched pair $i$, let $D_i = A(h_i^L) - A(h_i^R)$. With $P$ retained pairs, the aggregate magnitude and tie aware rank statistic are

$$G = \frac{1}{P} \sum_{i=1}^{P} |D_i|, \qquad \rho = \mathrm{corr}(\mathrm{rank}(a_i), \mathrm{rank}(|D_i|)) \,. \tag{27}$$

The active signature used by the rank statistic is

$$a_i = \tanh\left( \|m_i^L - m_i^R\|_2 + \frac{1}{2}\|u_i^L - u_i^R\|_2 + \frac{1}{4} \,\mathrm{mean}\, |b_i^L - b_i^R| \right). \tag{28}$$

Tied values receive their average rank, and $\rho$ is defined as zero when every $D_i$ is exactly zero. The area in Equation 26 is a bounded functional after state augmentation, but it is distinct from the binary hitting probability in Equation 2. Thus $G$ measures paired discrepancy for the declared empirical matching scheme rather than directly estimating the supremum $\Delta_B$.

For the controlled language-model panel, let $\theta_{\mathrm{LM}}(s)$ be the mean signed recovery-area difference for seed $s$ and let $\theta_{\mathrm{ref}}(s)$ be the frozen reference-panel mean. The continuous transfer statistic is

$$d_s = \theta_{\mathrm{LM}}(s) - \theta_{\mathrm{ref}}(s). \tag{29}$$

# B Experimental Design

## B.1 Finite Process and Observable-State Matching

The finite process uses an eight-dimensional logistic student $s_\theta(z) = \sigma(z^\top \theta)$ and a fixed logistic teacher. Features combine a scalar context, the student's sampled branch, their interaction, two smooth context transforms, and noise. The teacher parameter is $(0.35, 0.8, -0.2, 0.45, -0.5, 0.25, 0.1, -0.15)$ with temperature $0.9$. Every on-policy batch contains 16 trajectories. The student uses Adam with learning rate $0.003$, moments $(0.9, 0.99)$, numerical constant $10^{-8}$, multiplicative learning-rate decay $0.995$, and float64 arithmetic. Its persistent state also contains a tangent average with decay $0.9$, the most recent 16-example replay batch, occupancy and task counters, sampler position, and random tapes. Thus every field of Equation 1 has an executable realization.

Each training seed generated two histories from the same initialization and update rule but independent, side-indexed tapes from the same sampling law. The q01, q04, and q16 settings perform 1, 4, and 16 on-policy construction updates. For each seed and query setting, a SHA256 counter defines at most 1,664 candidate pairs. Before any recovery curve is evaluated, a selector visits candidates in increasing digest and candidate order and retains the first 416 pairs that match on seven declared coordinate families. The five vector families are logits on 32 fixed probes, divergence to the teacher, entropy, support overlap, and branch mass. The two count families are branch occupancy and task success at every construction prefix. Recovery outcomes, active signatures, $G$, and $\rho$ are unavailable to the selector.

Matching uses fixed numerical resolutions and outcome blind ordering. The primary pointwise resolutions for logits, divergence, entropy, overlap, and branch mass are respectively $0.15$, $0.03$, $0.03$, $0.15$, and $0.08$. The primary maximum count-rate difference is $0.25$ at every frozen prefix. Strict and loose selections repeat the ordering at half and twice the vector resolutions, with count-rate bounds $0.1875$ and $0.375$. We retain continuous results from all three selections. In the main census, the maximum absolute movement between adjacent selections is denoted $B_G$ for $G$ and $B_\rho$ for $\rho$, which report sensitivity to matching resolution.

Across the 12,480 primary pairs, median residuals were 0.0145 for logits, 0.00419 for teacher divergence, 0.00106 for entropy, zero for support overlap, 0.00353 for branch mass, and 0.125 for both count-rate coordinates. Recomputing every cell under the strict and loose selections changed $G$ by at most 2.73e-05 and $\rho$ by at most 0.0731 relative to an adjacent matching resolution. The artifact retains every coordinatewise median, upper quantile, maximum, and cellwise sensitivity value.

After selection, both members of a pair use a common future tape and the same teacher-query policy. A fixed 64-probe set is scored at recovery steps 0, 1, 2, 4, 8, and 16. Equations 26, 27, and 28 define the normalized recovery area, mean absolute paired difference $G$, active signature, and tie aware rank association. Each main cell contains $P = 416$ retained pairs. The continuous area and the binary hitting probability in Equation 2 are distinct recovery functionals, while exact full-state equality tests Proposition 1.

The full-state sufficiency control used the same seeds, query settings, and recovery scorer. Two independent execution paths were first generated and then every declared state carrier was set to the same canonical value before recovery. A clone control repeated the comparison from identical serialized state. These controls tested the pathwise premise of Proposition 1. Single-carrier interventions replaced optimizer, replay, teacher, or local-geometry state, while the natural-visitation and state-query families used common accounting for teacher queries, scored states, returned information, updates, and floating-point operations. Experience replay is a standard mechanism by which past samples affect later updates (Schaul et al., 2015), and target-network or replay state similarly extends the effective state of deep reinforcement-learning algorithms beyond their current outputs (Mnih et al., 2013).

### B.2 Exact Weight Fiber with Optimizer Memory

For hidden optimizer state $k$, future tape $j$, and initially incorrect evaluation state $e \in \mathcal{E}$, let $\tau_{\mathrm{rec}}^{k,j,e}$ be its exact recovery time. We estimate

$$\widehat{R}_B^{(k)} = \frac{1}{M|\mathcal{E}|} \sum_{j=1}^{M} \sum_{e \in \mathcal{E}} \mathbf{1}\{\tau_{\mathrm{rec}}^{k,j,e} \le B\}, \tag{30}$$

and report the finite subset diameter

$$\widehat{\Delta}_B^{\mathrm{subset}}(\theta_0) = \max_{1 \le k \le K} \widehat{R}_B^{(k)} - \min_{1 \le k \le K} \widehat{R}_B^{(k)} \le \Delta_B(o_{\theta_0}; \mathcal{Q}), \qquad o_{\theta_0} = \phi_\theta(x_k) = \theta_0. \tag{31}$$

The empirical subset diameter is the maximum minus minimum recovery probability among the eight hidden Adam states. It is a lower bound on the supremum over the full weight fiber. The base seed is the independent replication unit; common future tapes are paired repeated measurements. Reported intervals resample the eight base seeds. The causal reset sets every first and second moment to zero while preserving the shared step count. A second control uses SGD without momentum, which has no order dependent persistent state under zero learning rate.

### B.3 Nonzero Learning Rate Natural Histories

Signs are oriented from the selected history with larger first moment norm to the history with smaller norm, without using recovery outcomes. We report common tape signed contrasts, absolute contrasts calibrated by a tape sign flip null, exact sign flip inference over base seeds, and base seed bootstrap intervals. Future tapes remain paired measurements rather than independent replications.

The procedural setting uses eight base seeds and a nonzero learning rate of $5 \times 10^{-5}$ throughout warmup, construction, and recovery. Each base state branches into 128 histories that perform 12 updates. We observe restricted branch log probabilities on all 64 procedural states and answer log probabilities on every state and branch. An outcome blind rule searches all 8,128 candidate pairs and minimizes maximum function difference, followed by root mean square function difference, construction count difference, and candidate identifiers. There is no matching acceptance threshold, all eight selected pairs are retained, and each pair uses 64 common future tapes.

The Qwen3.5 GSM8K setting uses a 4B rank eight LoRA student and a 9B teacher (Cobbe et al., 2021; Yang et al., 2025). Both models use bfloat16. The teacher supplies a renormalized top 16 next token

distribution on student generated prefixes. Construction and recovery use ordinary nonzero learning rates of $10^{-4}$ and $5 \times 10^{-4}$. For each of eight declared base seeds, 32 candidate histories perform eight construction updates. The selector lexicographically orders current function $L_\infty$ distance on 16 fixed probes, RMS distance, probe top one disagreement, and candidate identifiers. Its input is the current function, with optimizer state and future recovery excluded and the acceptance threshold set to null. The recovery target is the first hash ordered held out problem answered correctly by the teacher and incorrectly by both selected students. This event is instantiated for seven of eight base seeds, with the eighth retained in coverage accounting. The seven pairs use 32 common future tapes and budgets 1, 4, and 8. We retain exact normalized answer correctness and the geometric mean probability of the reference solution tokens.

### B.4 Controlled Language-Model Stress Test

The controlled language-model study instantiates observable-state matching in named autoregressive systems. The decoder-only panel used DistilGPT2 and GPT2-medium, whose architecture descends from GPT-2 (Radford et al., 2019). The encoder-decoder panel used FLAN-T5-small and FLAN-T5-base, combining the text-to-text architecture of T5 with instruction tuning (Raffel et al., 2020; Chung et al., 2022). GSM8K supplied the public reasoning tasks (Cobbe et al., 2021), and the implementation used the Transformers model interface (Wolf et al., 2020). Prompt construction, data revision, tokenization, and greedy decoding were fixed across the panel. Comparability coverage and measured recovery are recorded separately: identities with the prescribed matched-state panel contribute a continuous outcome, and the remaining identities contribute to coverage accounting.

One identity is a model family, query setting, construction variant, and training seed. Each measured identity uses 40 outcome-blind selected comparisons. Equation 29 defines $d_s$ as the language-model mean signed recovery-area difference minus its frozen reference-panel mean. The complete language-model panel contained 270 identities. Continuous measurements were available for 178 identities, comprising all 90 decoder-only identities and 88 encoder-decoder identities. Two encoder-decoder identities remained in coverage accounting, and the 90 reasoning-anchor identities supply domain coverage outside the transfer comparison. Table 7 reports coverage and continuous outcomes together.

### B.5 Prospective New-Seed Validation

The theorem statements, analysis targets, and prospective validation protocol were frozen before generating new output. Six hashed seeds were disjoint from the ten main seeds and evaluated at q01, q04, and q16 with 64 matched pairs per cell, giving 18 cells. Signed bias, mean absolute error, and root mean squared error quantify stability on the new cells under this frozen specification.

The prospective cells were distributed across two independent accelerator hosts with source-identical code and one process per available device. Host type appears in the reproducibility record, while scientific aggregation uses the declared seed and query cells. Each host validated the protocol, prediction source, and source-code hashes before computation. The artifact archive contains the protocol, source, environment records, raw cell manifests, per-file hashes, seed derivation, aggregation code, and all compact scientific arrays used by the tables and figures. The aggregate verified the exact 18-cell universe and computed the declared continuous errors.

### B.6 Measured Outcomes

Every exact weight audit passed. Within each base seed, all eight complete weight hashes were identical, their maximum absolute parameter difference was zero, all eight Adam state hashes were distinct, and all histories shared one minibatch multiset and one future tape collection. Table 4 reports the 64 tape estimates. At budget one, the prespecified identity order minus reverse order contrast was $-0.042$ with base seed interval $[-0.099, -0.004]$ and two sided exact sign flip $p = 0.0156$. The split sample procedure selected two histories on 32 tapes and estimated separation 0.100 on the remaining 32, with interval [0.033,0.175]. The raw diameter was 0.104, its within tape permutation null mean was 0.016, and their calibrated difference was 0.088 with interval [0.029,0.155]. For the continuous recovery area, split sample separation was 0.026 with interval [0.015,0.036], and the calibrated diameter was 0.022 with interval [0.012,0.033]. Canonical Adam reset and SGD without

Table 4: Exact weight optimizer memory experiment with 64 common future tapes. Signed pair is identity order minus reverse order. Split uses 32 tapes to select two histories and the other 32 to estimate their separation. Null is the mean diameter under within tape label permutations, and calibrated subtracts this null mean. Intervals resample the eight base seeds.

| **Functional** | **Signed pair** | **Split** | **Raw $\widehat{\Delta}$** | **Null** | **Calibrated $\widehat{\Delta}$** |
|---|---|---|---|---|---|
| Exact, $B = 1$ | $-.042\,[-.099, -.004]$ | $.100\,[.033, .175]$ | .104 | .016 | $.088\,[.029, .155]$ |
| Exact, $B = 4$ | $-.0001\,[-.0004, 0]$ | $.039\,[0, .116]$ | .043 | .009 | $.034\,[0, .101]$ |
| Exact, $B = 16$ | $.002\,[-.0004, .007]$ | $-.0002\,[-.0007, 0]$ | .006 | .004 | $.001\,[0, .003]$ |
| Probability area | $-.015\,[-.031, -.001]$ | $.026\,[.015, .036]$ | .030 | .008 | $.022\,[.012, .033]$ |

Every base seed had one byte identical weight hash and eight distinct Adam state hashes.
Canonical Adam reset and stateless SGD controls reproduced exactly for every base seed.

Table 5: Continuous observable-state matching results and prospective new-seed validation. The new-seed columns use six unseen seeds and 64 matched pairs per cell.

| **Query** | **Main $G$** | **New $G$** | **$G$ MAE** | **Main $\rho$** | **New $\rho$** | **$\rho$ MAE** |
|---|---|---|---|---|---|---|
| q01 | 1.74e-04 | 1.61e-04 | 2.04e-05 | 0.196 | 0.141 | 0.0876 |
| q04 | 2.68e-04 | 2.87e-04 | 2.18e-05 | 0.091 | 0.132 | 0.0734 |
| q16 | 3.83e-04 | 4.08e-04 | 3.15e-05 | 3.93e-05 | -0.044 | 0.0503 |

momentum reproduced exactly for every base seed. Later budgets show the expected reduction as cumulative recovery reaches its ceiling.

In the procedural setting, the selected pair distance is about 15.5% of the candidate pair median and the average parameter relative L2 distance is $6.88 \times 10^{-4}$. At budget one, the tape calibrated optimizer main effect has mean magnitude 0.0570 with base seed interval [0.0055,0.1345], compared with a weights main effect of 0.0098. At budget four, the optimizer magnitude is 0.1461 with interval [0.0981,0.1872]. The continuous recovery area gives an optimizer magnitude of 0.0087 with interval [0.0051,0.0126].

Effect signs vary across base seeds in both settings, consistent with a fiber criterion that tracks recovery-law variation at each state. Together with the exact weight fiber, the two settings show that optimizer state remains active under ordinary nonzero learning rate training in controlled and mathematical reasoning tasks.

Across both settings, optimizer cross swaps retain more short horizon recovery variation than the corresponding weights main effects, while signed effects vary across base seeds. The joint result identifies state dependent optimizer relevance under ordinary training across moment norms, model scales, and task families.

Table 5 compares the original continuous measurements with the prospective validation. The main mean $G$ increased from 1.74e-04 at q01 to 3.83e-04 at q16. The new-seed means followed the same ordering, from 1.61e-04 to 4.08e-04. Across all 18 prospective cells, the frozen per-query predictions had a $G$ mean absolute error of 2.46e-05 and a root mean squared error of 2.74e-05. The external mean $\rho$ was 0.141 at q01, 0.132 at q04, and -0.044 at q16, while its overall prediction error was 0.0704 in mean absolute terms. The increasing magnitude trend persisted, while the q16 rank association changed sign. Recovery-gap scale and individual-history ordering are therefore distinct transfer targets.

The descriptive intervention census showed mechanism dependent signs. Table 6 reports the mean signed change in recovery area and the number of positive and negative seeds. Continued natural sampling was negative for every q01 seed but positive for every q04 and q16 seed. Fresh-root restart and ancestor-state queries had mixed signs. Increased rollout exposure and the hazard variant were negative in almost every tested cell. Directional targets were positive for all seeds, and top-k guidance and full-distribution guidance were also positive throughout. The latter two rows were numerically identical because the selected top-k payload was saturated in this finite process, so they represent one mechanism. These observations agree with Theorem 4: the value of direct access depends on the signed teacher response on the queried state distribution.

Table 6: Signed intervention effects. Each cell reports mean recovery-area change followed by positive and negative seed counts.

| Mechanism | q01 | q04 | q16 |
|---|---|---|---|
| Continued sampling | -7.75e-04 (0, 10) | 7.00e-04 (10, 0) | 5.35e-03 (10, 0) |
| Fresh-root restart | -3.07e-05 (4, 6) | -3.46e-05 (5, 5) | 9.26e-05 (6, 4) |
| Rollout exposure | -3.20e-04 (0, 10) | -3.31e-04 (0, 10) | -2.07e-04 (0, 10) |
| Hazard variant | -1.88e-04 (1, 9) | -5.56e-04 (0, 10) | -4.71e-04 (0, 10) |
| Directional target | 4.76e-04 (10, 0) | 1.33e-03 (10, 0) | 4.85e-03 (10, 0) |
| Top-k guidance | 1.44e-03 (10, 0) | 2.60e-03 (10, 0) | 4.90e-03 (10, 0) |
| Full guidance | 1.44e-03 (10, 0) | 2.60e-03 (10, 0) | 4.90e-03 (10, 0) |
| Ancestor-state query | 6.71e-05 (6, 4) | -1.75e-04 (3, 7) | 8.31e-05 (5, 5) |

The full-state controls completed 60 cells and produced zero recovery discrepancy in every cell. By contrast, $G$ was positive in all 30 observable-state matching cells. This combination supports the operational distinction behind the theory: the declared full state was sufficient for the tested process, while the reduced summary left a small but reproducible magnitude difference. The rank results identify state specific ordering across query settings. Fig. 4 shows the magnitude and rank measurements, and Fig. 5 relates the intervention signs to visitation and reliability.

All cells retain their original seed and query identities and enter the reported summaries. The result tables jointly show persistent recovery-gap magnitude after observable matching and state dependent ordering by the active signature. The complete cell universe connects these measurements directly to the fiber criterion.

# C Results and Analysis

## C.1 Magnitude Transfer and Ordering

The main census and the prospective validation separate two questions. As shown in Table 5, recovery-gap magnitude increased with the number of on-policy query opportunities in both the original and new-seed groups. One question concerns whether observationally matched histories retain a measurable difference in future recovery. The other asks whether a visible active signature orders those differences consistently across histories. The measurements support a transferable group-level magnitude trend at a small continuous scale, while the q16 sign change rules out the same conclusion for the proposed ordering statistic. The varying signed biases favor a query-dependent mechanism in which the opportunity process changes with q and hidden carriers accumulate at different rates.

Operationally, a magnitude forecast sizes a recovery study or compares query settings, while a ranking rule prioritizes individual states.

The prospective validation shows that a group-level effect scale can generalize more reliably than a state-level ordering. This distinction matters for deployment. A planner uses the typical recovery-gap magnitude to allocate monitoring budget, while a controller uses a stable ranking to choose individual states. Group-level sizing and state-level prioritization are therefore distinct prediction tasks, as reflected by the separation between $G$ and $\rho$.

The new-seed result also clarifies how to interpret the scalar recursion. Equation 20 predicts that opportunity depth can change the scale and sign structure of recovery drift, while the active signature tests a distinct observable coordinate. The observed growth of $G$ with q is consistent with stronger accumulation of history-dependent effects. The scalar theorem supplies the bistable phase portrait, and the measured neural process occupies its small-gap boundary case.

## C.2 Full-State Sufficiency and Intervention Mechanisms

The full-state controls provide the cleanest link between theory and measurement. Table 7 records zero discrepancy in every independent-path and clone control cell. This matches Proposition 1 when every variable read by the future transition is equalized. Under the declared scorer, tape, and

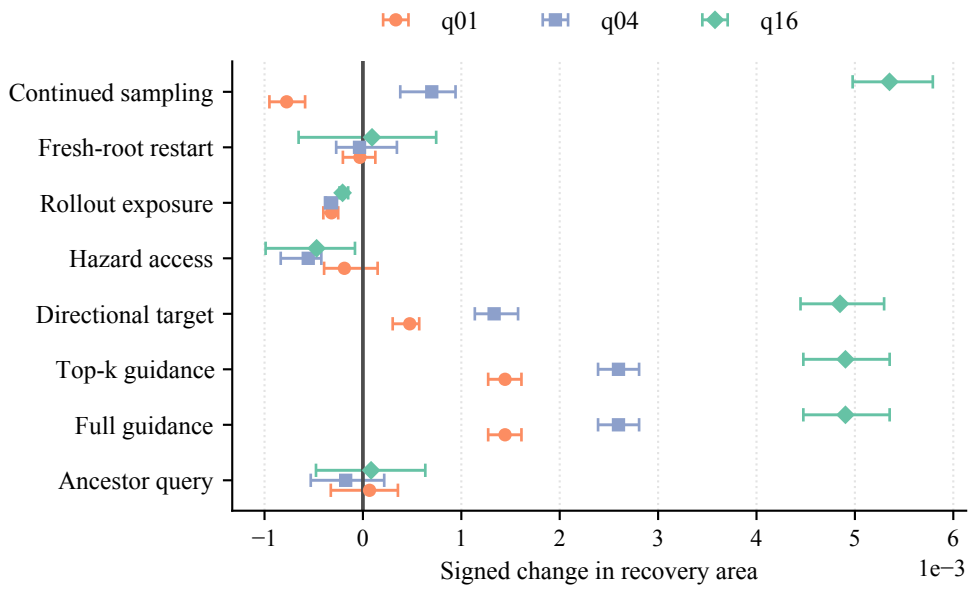


Figure 5. Signed intervention effects by mechanism and query setting. Points show seed means and bars show observed seed ranges. The zero reference distinguishes benefit from harm.

budget, canonicalizing all named carriers removes the discrepancy between the independent paths and supports the state decomposition exercised by the finite process.

The intervention atlas in Table 6 shows why one omitted-state label is insufficient. Directional targets and distributional guidance were positive across seeds, while increased rollout exposure and the hazard variant were negative. Ancestor-state queries had mixed signs. Common accounting holds queries, scored states, returned information, updates, and computation fixed across families, leaving the information entry point and its state-conditional sign as the varying mechanisms. This pattern follows the decomposition in Equation 10: direct access removes a visitation factor and multiplies the conditional teacher response.

The identical top-k and full-distribution rows provide a useful control on narrative multiplicity. In this process the top-k payload was saturated, so both methods returned the same effective information and represent one distributional-guidance result. The directional intervention remains distinct because it changes the target construction, and the ancestor query remains distinct because it changes the state at which information is requested. This grouping is consistent with recent OPD work that separates reliable token regions from outlier regions (Xing et al., 2026) and with methods that combine on-policy generation with off-policy reuse (Ko et al., 2024).

The sign reversals also separate visitation from accumulated update state. Continued sampling changed from unfavorable at q01 to favorable at larger q, whereas increased rollout exposure remained unfavorable. Their divergence identifies the ordering and context of feedback as active factors. The naturally matched atlas measures a collective carrier effect and motivates the exact construction, whose reset identifies optimizer memory under a controlled fiber. Replay, teacher, and tangent effects in the broader atlas retain their collective interpretation.

### C.3 Controlled Neural Panel

The language-model stress test separates measured continuity from comparability coverage. Decoder-only and encoder-decoder families produced continuous measurements for 178 identities. Table 7 reports pooled values by family and query setting across the three fixed construction variants. Decoder-only q01 had mean $d_s = 3.22\text{e-}04$ with standard deviation 2.60e-03, while the encoder-decoder mean was $-4.58\text{e-}04$ with standard deviation 2.68e-03. At q04 and q16, both family means were closer to zero and their ranges crossed zero. The sign varies with model family, query setting, and construction variant.

Table 7: Full-state controls and continuous language-model results. Language-model values pool three fixed construction variants.

| Family | Query | $n$ | Mean, SD | Range / control |
|---|---|---|---|---|
| Independent paths | all | 30 | $G_{\mathrm{CS}} = 0$ | Maximum $G_{\mathrm{CS}} = 0$ |
| Full-state clone | all | 30 | $G_{\mathrm{CS}} = 0$ | Maximum $G_{\mathrm{CS}} = 0$ |
| Decoder-only | q01 | 30 | 3.22e-04, 2.60e-03 | -4.86e-03 to 9.39e-03 |
| Decoder-only | q04 | 30 | 6.84e-06, 3.72e-05 | -7.51e-05 to 6.62e-05 |
| Decoder-only | q16 | 30 | 6.92e-06, 9.54e-05 | -2.13e-04 to 2.51e-04 |
| Encoder-decoder | q01 | 30 | -4.58e-04, 2.68e-03 | -5.50e-03 to 3.57e-03 |
| Encoder-decoder | q04 | 30 | 2.39e-05, 6.28e-04 | -1.17e-03 to 1.05e-03 |
| Encoder-decoder | q16 | 28 | 4.12e-06, 6.17e-05 | -1.03e-04 to 1.24e-04 |
| Reasoning-anchor | all | 0 | Coverage only | 90 declared identities |

Technical comparability defines the domain of the continuous recovery measurement. Table 7 records measured recovery for formed matched panels and coverage for the remaining identities, preserving equality of measured recovery and availability of a valid comparison as distinct events. The same distinction appears in recent empirical OPD analyses, where teacher-student compatibility and genuinely new teacher capability are separate conditions for useful training (Li et al., 2026).

### C.4 Operational Boundaries

Three operating boundaries follow from the measurements. First, magnitude calibration and state ordering are distinct targets. The active signature had positive association in some query settings and near-zero or negative association in others. It can track average recovery difficulty while assigning a different ordering to individual histories. This pattern is consistent with underdetermination: optimizer, replay, teacher, and local-geometry variables can produce similar current summaries and different signed future responses. State-level prioritization therefore calls for measurements that reveal the active carrier.

Second, teacher reliability determines the sign of increased access. The hazard and rollout families show unfavorable movement under additional exposure, while the directional and distributional families show favorable movement. Theorem 4 gives the corresponding design principle: estimate reliability on the queried state distribution and preserve the signed response. A practical diagnostic records the state variables consumed by the optimizer, separates natural and intervention distributions, and monitors signed effects together with query volume.

Third, comparability coverage defines the neural matched domain. Requiring agreement across many coordinates, prefixes, and task outcomes yields a narrower observation fiber. A lower-dimensional observable projection justified by the recovery event, larger prospective search capacity, or explicit reporting of the attained matched domain each gives a principled extension. In every case, the observation map is named before its fiber diameter is interpreted, preserving alignment with Theorem 1.

These boundaries define three validation targets. Magnitude calibration asks whether the distribution of $G$ transfers to new histories. Ranking calibration asks whether a chosen observable orders those histories. Comparability coverage asks which observation fibers are represented in a named model family. Treating the targets separately preserves the positive evidence carried by each aggregate and yields a clearer deployment decision.

These evaluation targets remain distinct: magnitude studies test group-level scale on new histories, ranking studies test prespecified state priorities, and coverage studies characterize the states admitted by matching. Progress on one target does not settle the others. Accordingly, each recovery estimate should specify its event, continuation policy, query budget, and matching variables. Comparisons across settings should make changes in these definitions explicit. The 90 reasoning-anchor identities in Table 7 contribute to coverage accounting but provide no measured recovery contrast; their absence from the measured panel is not evidence of zero discrepancy. Likewise, calibration of a group mean does not validate a rule for selecting individual states. These distinctions delimit the evidence available for each intended use.