HOW MUCH CAN THEIR SUCCESSFUL BEHAVIOR CHANGE BEFORE SUCCESS BREAKS?

Anonymous authors
Paper under double-blind review

ABSTRACT

An LLM agent can complete a task in different ways. Can one successful behavior be changed into another through valid edits while preserving reliable completion? We introduce interventional success regions: regions of reliably successful restartable checkpoints, linked by executable edits to their histories. An exact construction shows that success rates and specified local measurements can agree while successful connectivity differs. Budgeted discovery and independent path confirmation make this structure measurable. In ALFWorld and τ-bench retail, connectivity improves prediction on unseen task families, reducing Brier loss by 3.37%, with the largest gains under composed shifts. The average benefit extends to joint model and task-family holdout, with variation across models. As an application, a connectivity-gap-guided recovery curriculum improves shifted success by 5.12 percentage points over a difficulty-matched procedure. A separate evaluation holding the test-time selector fixed finds a 4.19-point advantage [95% CI: 1.88, 6.46], supporting a training benefit. Success-region structure thus provides information for both diagnosing agents and organizing recovery learning.

1 INTRODUCTION

An agent may complete a household task after locating an object immediately or after inspecting an extra location. Both histories succeed, but can one be changed into the other through valid intermediate histories that preserve reliable completion? This question matters when agents take detours, combine alternative decisions, or encounter several changes together.

Average success measures completion under nominal conditions; repeated trials measure consistency; perturbation tests measure local sensitivity. These measurements motivate a further question: **how far can successful behavior change through executable interventions before reliable completion breaks?** Reliability benchmarks and restart-based learning make this question operational (Yao et al., 2025; Lu et al., 2025). We study the connections among restartable checkpoints, each comprising an environment state, a consistent interaction history, and a remaining budget. Valid edits link checkpoints; continuation success determines which checkpoints belong to the success region.

Our contributions are:

1. **A structural foundation.** We define executable success regions and give an exact construction with identical labeled node values, degrees, anchor distances, and local widths, yet different successful connectivity.
2. **An executable measurement and diagnostic framework.** Budgeted discovery and independent confirmation make this structure measurable. Under fixed observations and specified local controls, connectivity improves prediction on unseen task families, including when the backbone is also held out.
3. **A recovery-learning application.** A connectivity-gap-guided curriculum improves shifted task success over a difficulty-matched procedure. A separate comparison with the test-time selector fixed supports a benefit from the trained acting model.

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

Figure 1 links the structural question to these diagnostic and learning uses. A successful trajectory records one sequence that worked; its success region describes how that behavior can change while reliable completion persists.

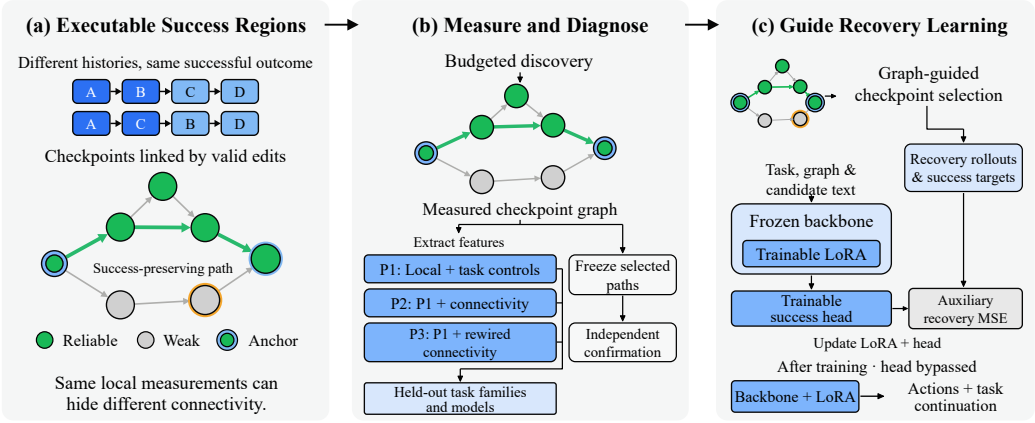

Figure 1: **From successful behavior to diagnosis and recovery learning.** (a) Valid edits connect restartable checkpoints; green nodes indicate reliable continuation. (b) Measured connectivity supports held-out prediction and independent confirmation of frozen paths. (c) Recovery targets train LoRA and a success head; the adapted backbone generates actions with the head bypassed. Graphs are conceptual.

2 From Successful Trajectories to Executable Success Regions

Executable checkpoints. For task x, a checkpoint $c = (s, h, b)$ records environment state s, consistent public history h, and remaining budget b. An agent configuration π includes its model, prompts, tools, decoding, and inference procedure. Its continuation value is $v_\pi(c; x) = \Pr(Y_x = 1 \mid \mathrm{Restart}(\pi, E_x, c))$. Restart restores the state and reconstructs input from the recorded history. An admissible checkpoint replays consistently and permits a successful continuation within budget; a verified successful rollout witnesses feasibility. A failed reference rollout leaves feasibility unresolved.

Success regions and local width. A declared edit registry defines a possibly directed graph $G_{\mathcal{I}} = (\mathcal{C}, \mathcal{E})$. Each edge is one validated history edit, executed through replay. Suppressing task and agent subscripts, the success region at threshold η and the local width after r edits are

$$\mathcal{S}_\eta = \{c : v(c) \geq \eta\}, \qquad W(r) = \mathbb{E}_{c\sim\mathcal{A},\, c'\sim Q_r(\cdot|c)}[v(c')], \tag{1}$$

where $\mathcal{A}$ samples shared anchors and Q_r samples r valid edits, allowing revisits. Width averages continuation success around these anchors; a successful source episode need not succeed on every restart.

Successful connectivity. Let $\mathcal{P}_L(A, B)$ contain admissible paths of at most L edits, and let $\mathcal{Q}$ be a declared bank of anchor pairs whose endpoints both meet the threshold. Define

$$b_L(A, B) = \max_{P\in\mathcal{P}_L(A,B)} \min_{c\in P} v(c), \qquad C = \frac{1}{|\mathcal{Q}|} \sum_{(A,B)\in\mathcal{Q}} \mathbf{1}\{b_L(A, B) \geq \eta\}. \tag{2}$$

The bottleneck is zero when no path exists. Connectivity measures the fraction of eligible pairs joined entirely through reliable checkpoints. Eligibility is reported separately; agent comparisons use pairs eligible for both.

Proposition 1. Two valued graphs can have identical labeled node values, per-node degrees, nearest-anchor distances, eligible pairs, nominal success, and repeated-run reliability, with equal anchor-averaged widths at every edit depth, yet different eight-edit connectivity.

Construction and proof idea. Use a sixteen-node cycle with four anchors and four attached leaves. Exchange the incident connections of nodes 6 and 10 while fixing every labeled value (Figure S1).

Degrees and nearest-anchor distances remain unchanged. Relabeling the second graph and using rotational symmetry proves equal anchor-averaged widths at every depth. Successful anchor groups nevertheless yield six versus four connected ordered pairs out of the same twelve. Appendix A gives the full proof and exact enumeration. Thus the specified local measurements leave connectivity undetermined; its predictive value is tested empirically below.

More explicitly, the successful anchor groups are $\{0, 4, 8\}, \{12\}$ in A and $\{0, 4\}, \{8, 12\}$ in B. Their within-group ordered pairs give $3 \cdot 2 = 6$ and $2 \cdot 1 + 2 \cdot 1 = 4$, respectively, with all connecting paths within eight edits. Both graphs have fourteen successful nodes out of twenty and widths $W(1) = 2/3$, $W(2) = 5/6$, and $W(3) = 7/9$. The example isolates the structural distinction without relying on different node values or different eligible-pair subsets.

The all-depth equality follows from the sampling symmetry. After relabeling B onto the original cycle, the two conditions fail at different pairs of the four gate positions $2, 6, 10, 14$, and at the same leaves. Rotation by four positions preserves the graph and uniform anchor distribution, so those gate positions have equal random-walk visitation probability at every depth. Their placement nevertheless partitions the successful anchors differently. Connectivity therefore captures how high-value checkpoints join, even when these averaged local measurements agree exactly.

3 Measuring Success-Region Structure

Discover, freeze, and confirm. Shared anchors have distinct semantic histories, incomplete goals, and matched progress. Allowed edits include action substitutions, information-gathering detours, bounded subroutine replacements, and verified independent-action reorderings. Editing an earlier action triggers replay and regeneration of affected observations; distinct histories remain distinct even at the same physical state. The evaluated agent generates all measured continuations; stronger-agent witnesses establish feasibility only.

Discovery prioritizes estimated bottleneck, using up to 32 validated nodes per search, eight continuations per node, and eight edits per bridge. This search cap differs from the size of the registered graph, whose leaves also enter feature calculation. The highest estimated-bottleneck path within the edit limit is frozen before confirmation. Proposal, replay, witness, and continuation costs enter the ledger (Appendix B).

For example, inserting an extra inspection into a household-task history changes the information available to the agent. Replay reconstructs the resulting observations, after which a continuation tests whether the task still succeeds. A sequence of such edits defines a route between histories. Each intermediate checkpoint must support reliable completion for that route to belong to the success region; successful endpoints alone do not establish a successful connection.

Measured connectivity features. Replace v in Equation 2 by node success rates $\widehat{v}$ on the registered graph to obtain $\widehat{b}$. For eligible ordered pairs $\mathcal{Q}$, the two features are

$$\widehat{C} = \frac{1}{|\mathcal{Q}|} \sum_{(A,B)\in\mathcal{Q}} \mathbf{1}\{\widehat{b}(A, B) \geq \eta\}, \qquad \widehat{B} = \frac{1}{|\mathcal{Q}|} \sum_{(A,B)\in\mathcal{Q}} \widehat{b}(A, B)\mathbf{1}\{\widehat{b}(A, B) \geq \eta\}. \tag{3}$$

We use $L = 8$ and $\eta = 0.8$. A below-threshold bottleneck contributes zero; both features are missing if there is no eligible pair. These graph estimates are distinct from independently certified bridge yield.

The features describe complementary properties: $\widehat{C}$ measures coverage by reliable routes, while $\widehat{B}$ also reflects their minimum node-success rates. Both depend on how node values are arranged across the measured edges.

Independent confirmation. For a frozen path of $m \leq 9$ vertices and J prespecified inspection sizes, fresh continuations give one-sided Clopper–Pearson lower bounds with per-bound error $0.05/(Jm)$. Certification requires all bounds to meet or exceed the threshold at the same inspection. Endpoint eligibility uses another sample bank. This allocation covers the registered path and inspections; additional paths require additional allocation. Failure to certify can reflect absent connectivity, incomplete discovery, or insufficient confirmation power.

Discovery supplies candidate paths; confirmation tests them with observations unused in selection. Replay failures, unresolved feasibility, no-path discovery, and failed confirmation remain separate outcomes. More confirmation can certify additional frozen paths; more discovery can change which paths are available.

The confirmation sweep follows each frozen path at six cumulative inspections: 32, 64, 128, 256, 512, and 1,024 samples per vertex. The full six-inspection allocation also applies at smaller caps (Appendix C).

The budget sweep separates discovery and confirmation effort: Reflect has higher point-estimate certified yield than Search, while every environment-specific paired interval includes zero (Appendix C). Execution-record checks pass 1,525/1,530 restorations and 401/408 first edge replays; 90/102 paths satisfy the complete restoration/first-replay criterion, including 53/59 originally certified paths. The 408 first edge attempts and seven diagnostic retries are counted separately (Appendix E).

4 Experimental Design

Environments and shifts. We study ALFWorld and τ-bench retail under initial-condition changes, interface/description changes, and their composition. Agent configurations include the backbone and execution harness. New-family diagnosis and joint backbone/family holdout use separate banks, each with 240 graphs; 236 eligible graphs contribute 2,832 common executions. All 48 executions from the four ineligible graphs remain recorded and are excluded equally across predictors.

The studies address complementary questions: diagnosis tests features on fixed observations; acquisition changes how observations are collected; recovery evaluates the actor after training. Acquisition uses 2,880 shifted target executions in a separate bank.

Fixed observations and controls. Predictors share checkpoint observations and target outcomes. P0 uses 25 inputs covering task/configuration controls, node summaries, and indicators. P1 adds 25 local-graph summaries, giving 50 inputs. P2 adds measured $(\widehat{C}, \widehat{B})$; P3 substitutes connectivity features from twenty rewired graphs while retaining all P1 inputs. All schemes use the same histogram-based gradient-boosting grid and family-based selection procedure (Appendix F).

The baseline includes nominal and repeated-run success, local widths, checkpoint-value summaries, task progress, and remaining budget. The added local-graph columns summarize degree, neighborhoods within one or two hops, and related availability indicators. This makes P1 a control for both observed success levels and specified local structure. Because P1/P2/P3 use the same continuation bank, their comparison measures feature value conditional on those observations. The acquisition study in Section 5 separately tests the value of changing which observations are collected.

Holdout and uncertainty. New-family predictors are fitted and selected only on f00–f09; five-model ensemble predictions evaluate f10–f19 without further fitting. Joint holdout excludes one backbone and one task-family group, giving thirty outer tests across six backbones and five groups. Brier loss averages squared probability error within family, then within environment, then equally across environments; the joint overall result also weights backbones equally. P3 averages seed-specific losses, with sample SD across twenty rewired versions. Paired 95% intervals resample families 10,000 times within environment with predictions fixed. Subgroup intervals are pointwise. Recovery intervals likewise retain observed trained models, seeds, and within-family cases. Appendices B, F, and G specify selection, pairing, missingness, and resampling.

Brier loss is $(\widehat{p} - y)^2$ for a predicted success probability $\widehat{p}$ and binary outcome y. Lower loss means more accurate probability prediction; a positive P1–P2 or P3–P2 difference favors measured connectivity. The reported relative reduction divides the P1–P2 difference by P1's Brier score.

Within the new-family training bank, selection uses five outer and three inner family folds. All tasks, configurations, shifts, and repetitions belonging to one family stay on the same side of each split. The five fitted outer-fold predictors form one frozen ensemble: their probabilities are averaged before evaluation, separately for each P3 rewiring seed. The f10–f19 outcomes enter neither fitting, tuning, nor method selection. This grouping makes the task family the unit of transfer; the five ensemble members are not five independent experimental replications.

Table 1: **Connectivity improves held-out success prediction on average.** P1/P2/P3 Brier scores (lower is better); P3 reports the mean with ± SD below, across twenty rewired variants. Paired-family 95% intervals appear below the gains; both use 10^{-3} Brier (positive favors P2). The two panels use separate banks of 2,832 eligible executions. Panel (b) has 480 executions per backbone except Haiku (432); its overall row weights backbones equally. Appendix F retains full precision and resampling details.

Model / cohort	P1	P2	P3 ± SD	P1–P2 [95% CI]	P3–P2 [95% CI]
(a) New task families					
New-family overall	0.1753	0.1694	0.1739 ±0.0006	+5.915 [4.204, 7.425]	+4.483 [3.294, 5.613]
(b) Joint backbone and task-family holdout					
Claude Opus 5	0.1547	0.1602	0.1554 ±0.0012	−5.557 [−10.560, −0.932]	−4.871 [−9.618, −0.356]
Claude Haiku 4.5	0.1926	0.1847	0.1914 ±0.0010	+7.900 [2.729, 13.189]	+6.672 [2.476, 11.020]
GPT-5	0.1719	0.1660	0.1704 ±0.0014	+5.834 [−0.033, 12.380]	+4.413 [−0.280, 9.470]
GPT-4o mini	0.2159	0.2089	0.2149 ±0.0014	+7.070 [1.127, 14.315]	+6.059 [1.607, 11.432]
DeepSeek-V4-Pro	0.1613	0.1545	0.1602 ±0.0009	+6.859 [1.577, 11.988]	+5.707 [1.876, 9.174]
DeepSeek-V4-Flash	0.1944	0.1882	0.1944 ±0.0015	+6.238 [−1.459, 13.486]	+6.220 [−0.452, 12.802]
Equal-backbone overall	0.1818	0.1771	0.1811 ±0.0007	+4.724 [2.803, 6.662]	+4.033 [2.227, 5.872]

5 Does Connectivity Predict OOD Success?

5.1 Added predictive value under two holdout settings

On unseen task families, measured connectivity reduces macro Brier from 0.175337 to 0.169422, a **3.37% relative reduction**. Both P1–P2 and P3–P2 intervals favor measured connectivity (Table 1). The rewired comparison asks whether the measured arrangement contributes beyond adding two graph-derived inputs to the same baseline.

The average benefit extends to joint backbone/family holdout: P2 reduces Brier by 0.004724 relative to P1 and 0.004033 relative to P3. This comparison contains 480 eligible executions per model except Haiku, which has 432. Table 1(b) gives model-level results; Appendix F.6 retains full precision and missing-cell details.

The local controls themselves are informative. In the new-family bank, moving from P0 to P1 reduces Brier by 0.001546 [0.000539, 0.002578] (Appendix F.3). Connectivity then adds the further improvement reported in Table 1. The P1–P2 and P3–P2 contrasts address two aspects of this result: improvement over the specified local baseline and improvement over a predictor with the same number of inputs using rewired connectivity. Their agreement supports the relevance of the measured connections to held-out prediction.

5.2 Where connectivity helps and where it fails

Families and composed shifts. P2 improves on P1 in 18/20 environment-specific test families and on P3 in 19/20. Gains concentrate under composed shifts (Figure 2): their gain exceeds the average single-shift gain by 0.016258 [0.011278, 0.020524]. Both composed groups and ALFWorld interface changes have positive pointwise intervals; other single-shift intervals cross zero.

Backbone variation. Haiku, GPT-4o mini, and DeepSeek-V4-Pro have positive intervals for both contrasts. Opus has negative intervals for both, while GPT-5 and DeepSeek-V4-Flash have intervals crossing zero. The equal-backbone average therefore captures a benefit whose magnitude and sign depend on the evaluated configuration.

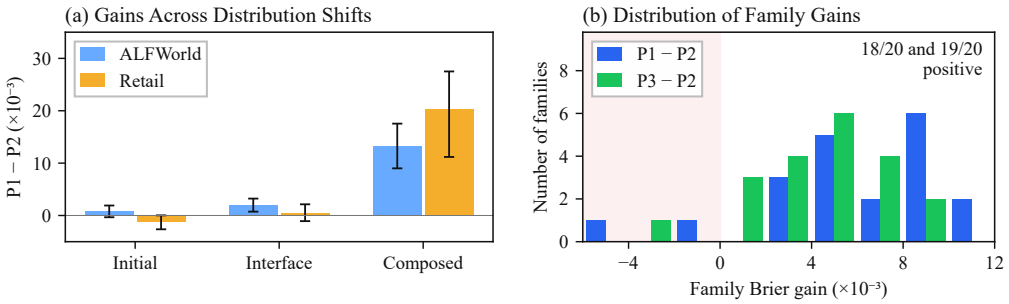

Figure 2: **Where connectivity improves diagnosis.** (a) Shift-specific P1–P2 gains with pointwise paired-family 95% intervals. (b) All twenty environment-specific family gains per contrast; red shading marks negative values. Units are 10^{-3} Brier; positive values favor P2. Predictions are fixed during resampling. Table 1 reports absolute scores and model-level contrasts; Appendix F gives full subgroup values.

Controls and outcome correspondence. The comparisons measure added value over the full P1 baseline. Six ALFWorld groups and five retail groups share all 25 local-graph features but differ in connectivity; other P1 inputs distinguish them. Appendix F.1 also identifies an earlier bank in which a two-hop feature determines connectivity. The empirical benefit is consequently evaluated relative to the declared representation and dataset. In R2, new graph measurements and outcomes give P1–P2/P3–P2 gains of +0.005771/+0.004744. Scoring the same predictions against the original 7G outcomes changes them to $-0.002398/-0.002130$, with both intervals below zero. This fixed-prediction stress comparison identifies correspondence between graph measurements and target executions as a condition for the observed benefit (Appendix F.5).

5.3 A COMPLEMENTARY USE: GUIDING ACQUISITION

Bridge-guided acquisition tests whether structure helps collect useful observations. Local, random-composed, and bridge-guided strategies share an initial bank, action allowances, graph-free predictor inputs, and family splits. At budget B, their macro Brier scores are 0.197, 0.194, and 0.191. The bridge-guided advantage over random acquisition is 0.003 [0.002612, 0.003384], with comparable realized action counts. The 0.5B/B/2B observations are cumulative snapshots of the same runs. Appendix D reports resource accounting and the stronger gains under composed shifts.

6 CAN CONNECTIVITY GUIDE RECOVERY TRAINING?

6.1 A CONNECTIVITY-GAP-GUIDED CURRICULUM

Select recovery practice. Each round uses a measured twenty-node graph and twelve ordered pairs of four successful anchors. For a shortest threshold-valid path of at most eight edges, the candidate is a minimum-success node on that path. Without such a path, the candidate is the lowest-numbered weak cycle node in the graph; all gap pairs share this fallback. Global scores gap pairs 0.8 and connected pairs 0.05, selecting the top four with deterministic pair-ID ties. Selected softmax weights over all twelve pairs are retained without renormalization. Rows sharing a node keep their separate rollout targets and weights. Appendix I gives the full Algorithm 1.

Breadth-first search visits neighbors in ascending node order and reconstructs paths through parent pointers, using only nodes with measured success at least 0.8. A gap pair's fallback is chosen from the whole cycle; its individual corridor is not consulted. Scores are rounded to four decimals before softmax and selection. These rules make candidate nomination and retained weights repro-

ducible, while each selected row preserves its pair identity, path status, and separate continuation observations.

Learn a representation for acting. Inputs encode task, measured graph, and candidate text. A sigmoid head reads the final-layer hidden vector at the last candidate-node-ID token. For nominal outcomes y_j and candidate success counts s_i from n_i recovery rollouts, the round objective is

$$\mathcal{L}_t = \frac{1}{2|\mathcal{N}_t|} \sum_{j \in \mathcal{N}_t} (\widehat{p}^{\,\mathrm{nom}}_{\theta,j} - y_j)^2 + \frac{1}{2} \sum_{i \in S_t} w_i \left(\widehat{p}^{\,\mathrm{rec}}_{\theta,i} - \frac{s_i}{n_i} \right)^2 . \tag{4}$$

The two loss coefficients are 0.5. Training freezes the base weights and updates q/k/v/o LoRA (rank 16, alpha 32, dropout 0.05) and the head, using AdamW at 10^{-4} with cosine scheduling for eight rounds of 500 steps. The backbone with LoRA generates actions; the head is bypassed at inference.

The predictions in Equation 4 are outputs of the learned success head; targets are observed nominal outcomes or complete-task recovery success fractions. Each candidate row retains four to eight recovery rollouts. The auxiliary objective updates the shared LoRA representation, whose behavioral effect is then assessed through generated actions and held-out task completion. When several selected rows refer to one checkpoint, their weights add while their samples and targets remain separate.

Evaluate task completion. Both pre- and post-training actors use the same post-training graph measurement, frozen before evaluation. After a failed nominal attempt, one recovery round tries at most four selected candidates: restore the checkpoint, apply one deterministic registered edit, then continue the complete task. Any successful attempt sets $y = 1$. Nominal and recovery attempts allow twenty and twelve interactions, respectively. Measurement costs are accounted for separately from evaluation interactions (Appendix G).

6.2 Matched curricula and a fixed test-time selector

We evaluate Qwen3.5-27B, Qwen3.5-9B, and Llama-3.1-8B-Instruct with five paired training seeds and 288 shifted cases per seed. Training families f00–f07 and test families f14–f19 are disjoint. Global and difficulty-matched recovery each specify 4,000 updates and a 92,000-interaction allowance. Both select only weak checkpoints and exactly one distinct checkpoint per round. Mean selected success is 0.449 versus 0.432; retained weight mass is 0.440706 versus 0.437586, a 0.708% difference (Appendix G.5).

The matched comparator nominates the weak cycle node with the lowest measured success rate and adds a small random offset to each pair's difficulty score. It provides a close comparison in selected difficulty, concentration, and total recovery-loss weight. Appendix G reports exact selection distributions and the scope of available cost records. Paired seed differences favor Global in 4/5, 4/5, and 5/5 runs across the three models; Appendix G.3 retains the seed-level exceptions.

With each method's own test selector, Global improves success by 4.375, 5.000, and 5.972 percentage points across the three models. The equal-model gain is **5.116 points [3.310, 6.875]**; all three model-level intervals are positive (Table 2(a)). The comparison supports the complete procedure, including its candidate nomination and learning inputs.

A separate crossed bank evaluates ten training conditions and three test selectors (129,600 outcomes; Table 2(b)). Holding the Global selector fixed, Global training gains **4.190 points [1.875, 6.458]** over difficulty-matched training and 9.005 [7.431, 10.394] over the untrained actor. Against matched training, the 27B and Llama intervals are positive; the 9B interval crosses zero. This supports trained-actor improvement with the test-time selection rule held fixed. The own-selector and fixed-selector banks support separate comparisons, not an additive decomposition.

Other recovery curricula. Under the same Global test selector, Local, auxiliary Difficulty, and Shuffled Global training reach 61.690%, 62.593%, and 61.505% success, respectively, versus Global's 69.005%. Their gains over the untrained actor are 1.690 [0.648, 2.616], 2.593 [0.718, 4.259], and 1.505 [−0.278, 3.356] percentage points, with paired-family 95% intervals. Local uses one-hop neighborhood scores; Shuffled Global permutes node values before path scoring. These comparisons evaluate alternative recovery curricula with test-time selection held fixed (Appendix G.6).

Table 2: **Recovery training improves shifted task success.** Success rates are percentages; Global-minus-comparator gains are percentage points with paired-family 95% intervals. Matched denotes difficulty-matched training. Each model/condition includes five paired seeds and 1,440 shifted cases (4,320 overall). Panels (a) and (b) use distinct outcome banks. Intervals condition on the observed training seeds; Appendix G provides seed dispersion, pairing, and variant definitions.

(a) Complete procedures: each curriculum uses its own test selector

Model	Matched	Global	Global–Matched [95% CI]
Qwen3.5-27B	65.625	70.000	+4.375 [2.083, 6.806]
Qwen3.5-9B	64.028	69.028	+5.000 [2.569, 7.292]
Llama-3.1-8B	62.014	67.986	+5.972 [2.500, 9.167]
Equal-model overall	63.889	69.005	+5.116 [3.310, 6.875]

(b) Acting-model comparison: fixed Global test selector

Model	Untrained	Matched	Global	Global–Matched [95% CI]	Global–Untrained [95% CI]
Qwen3.5-27B	60.972	65.833	70.000	+4.167 [1.181, 7.014]	+9.028 [5.764, 12.708]
Qwen3.5-9B	60.000	64.792	69.028	+4.236 [−0.486, 9.097]	+9.028 [5.972, 11.944]
Llama-3.1-8B	59.028	63.819	67.986	+4.167 [1.458, 6.736]	+8.958 [6.667, 11.458]
Equal-model overall	60.000	64.815	69.005	+4.190 [1.875, 6.458]	+9.005 [7.431, 10.394]

The primary bank evaluates the complete recovery procedure; the crossed bank evaluates the trained actor under a shared checkpoint rule. With the head bypassed for both actors, completion reflects actions generated by the adapted model. These tests connect auxiliary success prediction during training to observed recovery behavior.

Nominal-task performance. In the primary bank, Global and difficulty-matched training have identical nominal post-training means for each model. Each model contributes 480 nominal executions per condition across five paired seeds. Relative to the shared pre-training baseline, nominal success increases by 0.625, 0.625, and 0.417 percentage points for 27B, 9B, and Llama, respectively (Appendix G.3). The shifted-task gains thus accompany small observed improvements on nominal tasks.

6.3 Recovery targets and graph context

Under the fixed selector, nominal-only training has zero mean gain over the untrained actor, with interval $[-1.551, 1.829]$ points. Removing success counts lowers performance relative to Global by 4.514 [2.870, 6.134] points; removing graph context lowers it by 6.319 [4.213, 8.380]. These comparisons support recovery targets and graph context as useful components (Figure 3). Graph-context removal omits both node information and edges, and variants use different training records. On a shared candidate bank, a node-only rule reproduces Global's selected rows; Appendix G.5 separates this selection-stage equivalence from the full-procedure comparison.

7 Related Work and Discussion

Reliability and structural diagnosis. τ-bench measures completion and repeated-run reliability; broader work separates consistency, robustness, predictability, and safety (Yao et al., 2025; Rabanser et al., 2026). Recent connectivity studies examine parameter-space symmetries (Zhao et al., 2025; Theus et al., 2025) and paths among image inputs (Vrabel et al., 2025). Our paths link restartable agent checkpoints through executable edits and environment replay. Proposition 1 establishes a structural distinction under matched local measurements; held-out tests then measure its predictive value under specified controls.

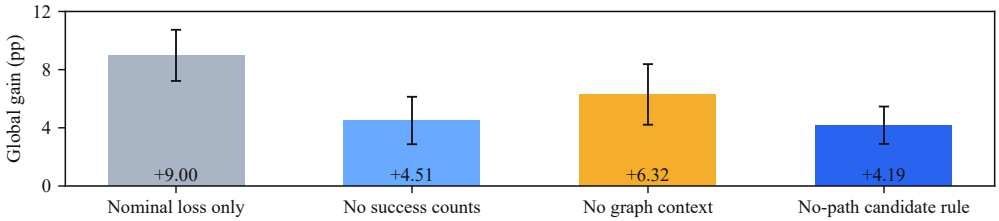

Figure 3: **Recovery targets and graph context support learning.** Global training minus each learning-input or candidate-rule variant under the fixed Global test selector, averaged equally over three models. Gains are percentage points with conditional paired-family 95% intervals; each condition has five seeds and 1,440 shifted cases per model. Graph-context removal omits both node information and edges; variants use different training records (Appendix G.7).

Refinement and restart curricula. AgentRefine studies feedback-based action refinement, while AgentNoiseBench tests tool-using agents under controlled user and tool noise (Fu et al., 2025; Wang et al., 2026). Intelligent Go-Explore uses foundation models to guide archiving, returning to, and exploring promising states (Lu et al., 2025). Our curriculum refers to bounded-path gaps among successful anchors. Difficulty-matched and fixed-selector comparisons evaluate the resulting recovery procedure and the adapted acting model.

Scope and interpretation. The evidence connects an exact structural separation, incremental held-out prediction, and an effective recovery procedure. The largest diagnostic gains under composed shifts motivate measuring relationships among alternative behaviors. Model heterogeneity and R2 identify conditions that affect transfer. Results concern registered edits, finite measured graphs, two environments, and the evaluated model banks; family intervals condition on recorded predictions or trained models. Candidate nomination, measured data, and graph context jointly define the recovery procedure. Appendices G–I report implementation correspondence and component comparisons, supporting future tests of individual structural interventions.

Behavioral flatness: flatness for agents. Reward flatness in policy-parameter space is linked to action robustness in reinforcement learning (Lee & Yoon, 2025). We develop an analogous behavioral view with the agent fixed: starting from a checkpoint $c^\star$ on a successful trajectory $\tau^\star$, does completion remain likely after executable edits? Let $\mathcal{N}_r(\tau^\star; c^\star)$ be the trajectory distribution induced by r registered edits, replay, and continuation by the fixed agent. Define behavioral flatness as $F_\pi(c^\star; r) = \Pr_{\tilde{\tau}\sim\mathcal{N}_r}[Y_x(\tilde{\tau}) = 1]$, where Y_x indicates task success. Equivalently, $F_\pi(c^\star; r) = \mathbb{E}_{c'\sim Q_r(\cdot|c^\star)} v_\pi(c'; x)$. The neighborhood is specified by the edit registry, sampling rule, and execution budget. We call the behavior flat over a declared range of edit depths when $F_\pi(c^\star; r) \geq \eta$ throughout that range.

Local width is its anchor average: $W(r) = \mathbb{E}_{c\sim\mathcal{A}} F_\pi(c; r)$. Proposition 1 shows that identical average flatness profiles can coexist with different successful connectivity. Behavioral flatness describes tolerance to sampled changes; connectivity further asks whether successful alternatives can be joined through reliable intermediate checkpoints. Together, they characterize the extent and organization of behavioral flexibility.

8 Conclusion

Executable success regions characterize how successful behavior can change while preserving reliable completion. Their connectivity distinguishes graphs with matched local measurements, improves held-out diagnosis in the evaluated settings, and guides recovery learning with a benefit under fixed test selection. The framework links behavioral flexibility to practical choices about what to measure, which failures to anticipate, and where to practice recovery. Understanding how much successful behavior can change calls for measuring both tolerance to sampled edits and the routes connecting successful alternatives. Executable success regions provide a common framework for studying these complementary aspects of agent reliability through registered interventions and independently testable paths.

GENERATIVE AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, language polishing, and LaTeX typesetting.

REFERENCES

Dayuan Fu, Keqing He, Yejie Wang, Wentao Hong, Zhuoma GongQue, Weihao Zeng, Wei Wang, Jingang Wang, Xunliang Cai, and Weiran Xu. AgentRefine: Enhancing agent generalization through refinement tuning. In *International Conference on Learning Representations*, pp. 65185–65204, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a3cc50126338b175e56bb3cad134db0b-Abstract-Conference.html.

Hyun Kyu Lee and Sung Whan Yoon. Flat reward in policy parameter space implies robust reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3448058d44f133839042295f79a9a958-Abstract-Conference.html.

Cong Lu, Shengran Hu, and Jeff Clune. Intelligent Go-Explore: Standing on the shoulders of giant foundation models. In *International Conference on Learning Representations*, pp. 21276–21301, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/369a30aac2765950865efd318cef7f76-Abstract-Conference.html.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of AI agent reliability. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66364.

Alexander Theus, Alessandro Cabodi, Sotiris Anagnostidis, Antonio Orvieto, Sidak Pal Singh, and Valentina Boeva. Generalized linear mode connectivity for transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4546. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c6eadcc507edc04c1baf00f05cd18b1a-Abstract-Conference.html.

Jakub Vrabel, Ori Shem-Ur, Yaron Oz, and David Krueger. Input space mode connectivity in deep neural networks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d75660d6eb0ce31360c768fef85301dd-Abstract-Conference.html.

Ruipeng Wang, Yuxin Chen, Yukai Wang, Chang Wu, Junfeng Fang, Xiaodong Cai, Qi Gu, Hui Su, An Zhang, Xiang Wang, Xunliang Cai, and Tat-Seng Chua. AgentNoiseBench: Benchmarking robustness of tool-using LLM agents under noisy condition. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64355.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, pp. 9965–10017, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding mode connectivity via parameter space symmetry. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77441–77460. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhao25i.html.