SELECTION IS NOT EVALUATION: INDEPENDENT CONTINUATIONS FOR RUNTIME INTERVENTION IN LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting a runtime intervention and evaluating it on the same stochastic continuation credits favorable branch noise. We generate two independently seeded continuations for every action from a shared prefix. The same draws separate the action chosen from the outcome that scores it and, before any controller is fitted, read the room a decision point and a menu leave. On 384 identity-disjoint ALFWorld tasks the same-draw maximum overstates cross-draw value by 15.36 percentage points (floorplan-bootstrap 95% interval [12.46, 18.30]), and does not exceed a within-task exchangeable-label reference of 17.26 points on eligible prefixes. On 593 further tasks run under both rules with one menu, moving the decision from a scheduled early step to the first public failure signal turns a negative reading, 52.50% for a per-task selector given one independent draw of every action against 53.33% for the best arm in hindsight, into a positive one, 50.80% against 49.09%. A controller that chooses its action on one development draw and values it on another captures 79 percent of the room so created on triggered prefixes and leads every matched comparator, by 1.26 points over continuation on all planned tasks and by 1.94 over the best fixed repair; refitted on development records from the scheduled rule, it leads every comparator on that panel too.

1 INTRODUCTION

A language agent can be warned, asked to replan, or rolled back when a trajectory appears to be failing. None of these repairs is free: a warning can displace useful context, a new plan can abandon partial progress, and a rollback can undo a correct action. A runtime controller must therefore answer a treatment-choice question: which action, including no intervention, has the highest expected utility from the current prefix?

The evaluation is easy to bias. Suppose an experiment restores one prefix, samples several intervention branches, and then keeps the branch with the best observed outcome. Scoring that choice on the same branches measures a realized maximum and credits the selection rule for favorable continuation noise. The problem is acute in binary-success environments, where one lucky branch turns an otherwise tied menu into an apparent rescue, and repeating the task under a new seed does not help if the repeated outcome again both chooses and scores.

This distinction is separate from two established ideas: outcome regression and direct treatment-effect regression are alternative target parameterizations (Künzel et al., 2019), and separating selection from evaluation is a general remedy for maximization bias and adaptive reuse (van Hasselt, 2010; Dwork et al., 2015). We apply both to intervention policies for language agents. The closest same-prefix study evaluates its main controller with one realized branch per action (Zhang et al., 2026).

For each action we generate two independently seeded continuations from a shared prefix. We evaluate any policy fitted on separate development tasks on both draws, and for an outcome-informed oracle diagnostic we score the action selected on one draw against the other. This yields a symmetric, pathwise nonnegative gap between the observed maximum and the independent-draw value of the selected action. The gap measures how much a same-outcome analysis can overstate deployable utility on the sampled task frame. It does not establish that any learned intervention helps.

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.

The same design answers a second question that a single draw cannot. Give an oracle one complete independent draw of every action and score it on another. The resulting value is what a per-task selector reaches from outcomes alone at this draw count. When it falls below the best arm in hindsight, the per-task outcome signal carries less than the menu's average effect, and the binding constraint is the design rather than the learner. The reading is a measurement, not a ceiling. A per-task selector is a per-unit estimate, and a learner that pools across tasks is a shrinkage estimator, which can dominate per-unit estimates when the units are many and each is noisy (Efron & Morris, 1975). Section 6 reports by how much under each rule. Our first confirmation, with four fixed messages at a hash-assigned early step, is the case where the reading is negative. Our second confirmation changes two design variables and leaves the estimand, the protocol, and the learner family intact: the decision is taken when the public prefix first shows an invalid action, an inadmissible action, or an unchanged observation, and the menu separates repair depth from repair content by adding a three-decision restoration and a relocation message. Its tasks are run under both rules with that one menu, so the decision point is the only thing that moves between the two readings.

A third use of the repeated branches is internal to the controller. A deployed rule also maximizes, over estimated benefits rather than over outcomes, so the score its margin test sees is itself a realized maximum. We fit one signed-benefit forest per development draw and choose the action on one draw while valuing it on the other, which moves the selection-versus-evaluation split from the evaluator into the controller at no extra data cost.

Across the controller comparison we hold the development records, prefix features, four whole-task folds, 200-tree capacity, margin grid, and action menu fixed. One forest predicts arm outcomes; a second predicts signed benefits relative to continuation,

$$\Delta_a(x) = \mathbb{E}[Y_a - Y_0 \mid X = x], \qquad a \neq 0. \tag{1}$$

Both can choose CONTINUE. The candidate set also contains the development-best fixed arm, the strongest runnable arm-outcome policy, a comparison-only learner, and a failure-risk detector. We committed every prediction before any confirmation outcome existed.

The first confirmation panel contains 384 ALFWorld tasks balanced over six task types and grouped by 96 floorplans, whose goal directories are absent from the supplied exposure packet, plus 57 reserved ScienceWorld variations as a boundary. The second adds 593 further identity-disjoint ALFWorld tasks. We report every frozen policy, costs, early-terminal tasks, and the adverse historical evidence.

We make three contributions. We give a repeated-branch protocol and a simple diagnostic that prevents a realized action maximum from being reported as policy value, together with a reference distribution that says how much of the diagnostic is mechanical. We turn the same protocol into a design instrument that measures how much per-task signal a decision point and a menu carry before any controller is fitted, and we show what a pooled learner adds to that reading. We provide a controlled target-parameterization comparison, extend it with cross-draw selection inside the controller, and release a pre-outcome configuration, prediction freeze, bounded compute ledger, full result census, and scripts that regenerate every table and figure. The empirical conclusions are stated in Section 5; they are not widened beyond the frozen tests.

2 EVALUATING A RUNTIME INTERVENTION

2.1 POLICY VALUE AFTER A SHARED PREFIX

Let X_i be the observable prefix for task i at its assigned checkpoint and let $E_i = 1$ when the episode ends before that checkpoint. The action set is $\mathcal{A} = \{0, \dots, K\}$, where 0 means continue without an injected message. A stochastic continuation under action a produces binary task success $Y_{ia} \in \{0, 1\}$ with conditional mean $\mu_a(x) = \mathbb{E}[Y_{ia} \mid X_i = x]$. A prefix-only policy $\pi(a \mid x)$ acts only on eligible prefixes, so the reported task-level value is a mixture,

$$V(\pi) = \mathbb{P}(E = 1)\,\mathbb{E}\big[Y^{\text{early}} \mid E = 1\big] + \mathbb{P}(E = 0)\,\mathbb{E}_{X \mid E=0}\left[\sum_{a \in \mathcal{A}} \pi(a \mid X)\mu_a(X)\right]. \tag{2}$$

Early-terminal tasks contribute their observed terminal outcome to every policy, so they cancel from every contrast and enter only the denominator. The expectation includes continuation randomness. A

single fresh draw per task is already an unbiased estimate of $V(\pi)$ for a policy committed before that draw; the bias studied next arises only when the action is chosen using the outcome that then scores it. A controller that always intervenes can have lower value than continuation even when it correctly detects a difficult prefix. Detection quality and treatment choice are different estimands (Vasudev et al., 2026).

Our simulated environments permit a repeated-branch design, and we build the estimand on it. From the same serialized state, we generate two draws $Y_{ia}^{(0)}$ and $Y_{ia}^{(1)}$ for every action. The request seeds are independent deterministic hashes of task, draw, action, and request index. Full-state hashes validate restoration before the first suffix call. The actor sees only the common public prefix, the assigned message, and subsequent observations.

2.2 The same-draw optimism diagnostic

Define $\widehat{a}_i^{(r)} = \arg\max_a Y_{ia}^{(r)}$, resolving ties by the lowest action index, so continuation wins every tie it enters and the lowest-indexed successful intervention wins otherwise. A same-draw analysis scores this outcome-informed choice as $\max_a Y_{ia}^{(r)}$, whereas cross-draw evaluation scores the selected action on the other draw. We use the symmetric gap

$$G_i = \frac{1}{2}\left[\max_a Y_{ia}^{(0)} - Y_{i,\widehat{a}_i^{(1)}}^{(0)} + \max_a Y_{ia}^{(1)} - Y_{i,\widehat{a}_i^{(0)}}^{(1)}\right]. \tag{3}$$

Proposition 1. $G_i \geq 0$ for every realized pair of action-outcome vectors. If the two action-outcome vectors are independent and identically distributed conditional on X_i, then $\mathbb{E}[G_i \mid X_i]$ equals the expected same-draw maximum minus the independent-draw value of the action chosen on that draw.

The pathwise statement follows because each maximum is at least the outcome of the action selected on the other draw; it needs no distributional assumption. Conditional independence is what factors the cross-draw term into the value of the selection rule on a fresh continuation, and identical distribution is what makes the two orderings interchangeable. Without independence the identity fails: two perfectly dependent draws give $G_i \equiv 0$ while the expected same-draw maximum still exceeds the independent-draw value whenever continuation outcomes are stochastic. With $R > 2$ draws the same statement holds for the average of G over ordered draw pairs. A positive G_i can arise from unstable binary outcomes even when all actions share a conditional mean. Because $G_i \geq 0$ pathwise, a sign-flip null is degenerate for it: every group sum is non-negative, so the test can only reject the sharp null of deterministic continuations. We therefore treat G as an estimation diagnostic and read it against a within-task reference that permutes the action and draw labels of the recorded outcomes, fixing the noise level and removing action structure. Appendix A gives the derivation and tie cases.

Figure 1 shows the information boundary, within which frozen policies use the prefix only. The outcome-informed selector exists solely to quantify same-draw optimism; it is not a deployable method or a training label for the confirmation policies.

3 Controllers with a continuation option

3.1 Matched target parameterizations

The repeated arm-outcome learner fits a multi-output random forest (Breiman, 2001) to

$$\bar{Y}_{ia} = \frac{Y_{ia}^{(0)} + Y_{ia}^{(1)}}{2}, \qquad a \in \mathcal{A}. \tag{4}$$

At prediction time it compares the estimated outcome mean of every action. The direct learner uses the same records but fits three outputs,

$$\bar{D}_{ia} = \frac{1}{2}\sum_{r=0}^{1}\left(Y_{ia}^{(r)} - Y_{i0}^{(r)}\right) = \bar{Y}_{ia} - \bar{Y}_{i0}, \qquad a \neq 0. \tag{5}$$

Its continuation score is fixed at zero. Thus the two learners differ in the response target while retaining the same action menu and predictor class. The comparison does not assume that direct effect

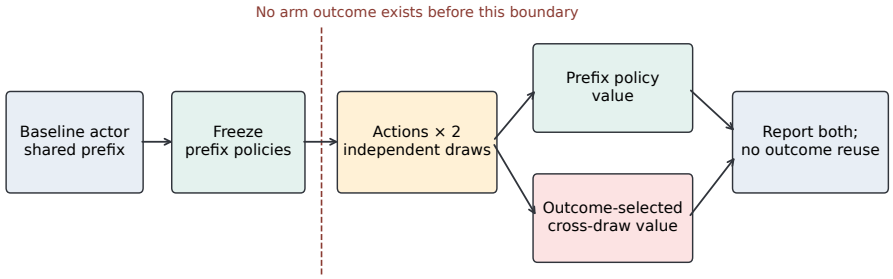

Figure 1: Repeated-branch confirmation. A baseline actor creates a checkpoint, at a hash-assigned step in the first confirmation and at the first public failure signal in the second. Every eligible prefix is restored for every action in the frozen menu and two independently seeded draws. Prefix-only controller predictions are frozen before any of these arm outcomes. Outcome-informed action choice on draw 0 is scored on draw 1 and vice versa.

regression is uniformly preferable; such an ordering is absent even in standard heterogeneous-effect settings (Künzel et al., 2019).

Both learners receive public information available at the fork: environment identifier, checkpoint index, counts of invalid or inadmissible actions, reward and usage summaries, action repetition signals, recent observations and actions, and recent prefix text. Hidden state hashes and reversibility metadata are excluded. Numeric fields are vectorized directly. Recent text uses a 512-term TF–IDF vocabulary fitted within each training fold and at most 16 truncated-SVD components.

Model selection uses four shuffled folds that keep every record of a task together, with seed 260908. Each candidate has 200 trees, minimum leaf size 5 or 10, and a continuation margin in $\{0.00, 0.05, 0.10\}$. Vocabulary and projection fitting occur inside each fold. Out-of-fold utility selects the configuration, and lower intervention rate breaks utility ties. The final forests are refit on all admitted development tasks. Source and target hashes, fold identities, feature schemas, library versions, and every candidate score are stored in the model manifests.

3.2 Cross-draw selection inside the controller

A deployed controller also maximizes. It compares estimates $\widehat{\Delta}_a(x)$ and fires the arm attaining their maximum, so the score that the margin test sees is a realized maximum over noisy estimates and inherits the bias of Section 2 at the estimation level rather than the outcome level. The repeated-branch design already supplies the remedy it needs. We fit one signed-benefit forest per development draw, $\widehat{\Delta}^{(0)}$ and $\widehat{\Delta}^{(1)}$, on identical features, folds, and capacity. The arm is chosen on one draw and valued on the other,

$$\widehat{a}^{(r)}(x) = \arg\max_a \widehat{\Delta}^{(r)}_a(x), \quad V^{(r)}(x) = \widehat{\Delta}^{(1-r)}_{\widehat{a}^{(r)}(x)}(x), \quad S(x) = \begin{cases} \frac{1}{2}\left[V^{(0)}(x) + V^{(1)}(x)\right] & \widehat{a}^{(0)}(x) = \widehat{a}^{(1)}(x), \\ \max_r V^{(r)}(x) & \text{otherwise,} \end{cases} \tag{6}$$

and the controller fires the arm attaining $S(x)$ when $S(x)$ exceeds the continuation margin. Two entries join the shared grid: whether to require $\widehat{a}^{(0)}(x) = \widehat{a}^{(1)}(x)$, which declines to act when the draws disagree about which repair the prefix calls for, and whether to restrict the candidates to the two arms with the highest mean benefit inside the training fold. Both are chosen by development out-of-fold utility before any confirmation outcome. Equation 6 is Equation 3 moved from the evaluator to the controller: the same symmetric construction applied to estimates rather than outcomes, and the same requirement that the quantity used to choose never be the quantity used to value.

Model selection faces the same problem once more, because the grid is scored and then maximized on the same folds. We therefore also report a nested out-of-fold utility that selects inside each training fold and scores on the held-out fold. Ties in that utility are broken by how often the nested inner folds made the same choice, the selection-frequency criterion of Meinshausen & Bühlmann (2010), rather than by the firing rate.

Table 1: Controlled comparison and frozen action menu. Arms $a = 0$ to $a = 3$ are shared by both confirmations.

Component	Arm-outcome forest	Direct signed-benefit forest
Development rows	Same two-draw task blocks	Same two-draw task blocks
Features and text	Shared prefix transform	Shared prefix transform
Folds and capacity	4 whole-task folds; 200 trees	4 whole-task folds; 200 trees
Leaf and margin grid	$\{5, 10\}$; $\{0.00, 0.05, 0.10\}$	$\{5, 10\}$; $\{0.00, 0.05, 0.10\}$
Response	$(\bar{Y}_0, \ldots, \bar{Y}_3)$	$(\bar{D}_1, \bar{D}_2, \bar{D}_3)$
No-intervention action	$a = 0$	zero-benefit reference
$a = 0$	Continue with no injected message	
$a = 1$	Check goal, observation, and previous action	
$a = 2$	Construct and execute a fresh concise plan	
$a = 3$	Undo the previous action, then replan; matched decision budget	
$a = 4, 5$	Second confirmation only: undo three decisions, then replan; relocate the goal	

3.3 Fallbacks and the comparator

Continue always chooses $a = 0$ and Best fixed always chooses the development-best arm, which is replanning ($a = 2$) in the first confirmation and the relocation message ($a = 5$) in the second. Direct uses the signed-benefit learner and Arm outcome the two-draw outcome forest; this pair is the sample-matched target-parameterization contrast, because both fit the same two-draw rows. Matched comparator is instead the strongest runnable member of the existing arm-outcome family by development out-of-fold utility, which is Single-1, a forest fitted to one predesignated draw; comparing Direct against it therefore measures the target parameterization together with the draw count, and is reported as the strongest available benchmark rather than as the clean parameterization test. The second confirmation adds Cross-draw and two learners that instantiate the closest external objectives inside this interface: a comparison-only forest fitted to the sign of $Y_a - Y_0$, and a failure-risk detector that gates a fixed arm on predicted continuation failure.

The operational candidate set is {Continue, Best fixed, Direct, Matched comparator}. Its development-only selection uses the same utility and tie rule. The selected policy is reported as Safe selected even when it is a baseline or an existing learner. In both environments the frozen selection chose Single-1. Calling this rule safe means only that it can decline intervention and was selected without confirmation outcomes; it is not a safety guarantee.

4 Independent confirmation design

We design two frozen confirmations to answer four questions in the order the results report them: (1) how much does same-draw action selection overstate independent-draw value, and how much of that is continuation noise; (2) does a matched controller improve on continuation and on the best fixed repair at a scheduled checkpoint with four fixed messages; (3) what do the same branches say that decision point and that menu could have delivered; and (4) does moving the decision to the first public failure signal, separating repair depth from repair content, and decoupling the controller's own choice from its own valuation change the answer to (2).

4.1 Actor and interventions

The actor is Qwen3-8B at revision `b968826d9c46dd6066d109eabc6255188de91218` (Yang et al., 2025) under a grounded ReAct-style prompt (Yao et al., 2023): one short thought followed by an exact valid action, thinking mode disabled, temperature 0.7, top-p 0.9, at most 256 output tokens, and a 32,768-token context. Serving uses deterministic-inference mode in a content-addressed SGLang container, which fixes the sampling implementation for the study without claiming invariance across serving systems (Bjarnason et al., 2026).

ALFWorld episodes have a 50-action limit and scheduled checkpoints at step 4 or 8 (Shridhar et al., 2021). ScienceWorld episodes have a 100-action limit and checkpoints at step 8 or 16 (Wang et al., 2022). A hash of each task identity assigns the checkpoint before generation. Action 3 restores the

state before the last executed action and reduces the suffix limit by one. The other arms restore the checkpoint state. This gives every arm the same total decision budget. The intervention message is inserted once at the restored step and remains in history; no controller sees future state or outcome fields.

The second confirmation keeps this actor, seeding, serving image, and restoration contract and changes only the decision rule and the menu. Its checkpoint is the first step between 2 and 24 at which the public prefix shows an invalid action, an inadmissible action, or an observation identical to the preceding one; a task with no such step has no intervention decision. Its menu keeps arms 0–3 verbatim and adds an arm that reverts the three most recent decisions before replanning and an arm that asks the actor to name the objects still required and the places not yet inspected. Reverting k decisions reduces the suffix limit by k, so every arm still costs the same total number of decisions. Its 593 identities are run under both the scheduled and the event rule, which makes the rule contrast paired at the task level.

Each task first receives one baseline episode. A task that terminates before its checkpoint has no intervention decision; its terminal success is assigned to every policy and its contrasts are zero. For every eligible task the continuation arms are generated fresh rather than copied from the baseline suffix, because that trajectory helped define the prefix.

4.2 Panels and exposure boundary

The primary panel has 384 ALFWorld tasks, 64 from each of the benchmark's six task types, drawn from the training source because it supplies enough identity-disjoint instances for a powered confirmation. Selection excluded 394 goal directories appearing in 59 supplied prior configurations or outcome records; the remainder were ranked by a fixed hash seed and form 96 floorplan groups. This is a fresh study panel relative to the supplied research archive, not a claim that the instances never appeared in pretraining. Development results favored ALFWorld, so the domain was fixed before sampling and before predictions were frozen. The secondary panel has 57 ScienceWorld official-test variations from a previously reserved frame, spanning 25 task families; no ScienceWorld outcome changes the primary family or controller tuning. The second study draws 593 confirmation identities and a disjoint development panel from the same inventory after also removing every goal directory used by the first panel. Appendix D gives the construction and exclusions.

The panel configuration and hypotheses were committed before baseline generation. After all baselines, the feature transforms and one-hot policy choices for 406 eligible prefixes were hashed and committed before any arm outcome. The freeze contains 355 ALFWorld and 51 ScienceWorld predictions. Appendix D gives the commit and SHA-256 lineage.

4.3 Hypotheses, uncertainty, and stopping

The first confirmation's ALFWorld family contains mean same-draw optimism $\mathbb{E}[G]$, Direct minus Continue, and Direct minus Matched comparator; the second confirmation's family contains Cross-draw minus Continue and Cross-draw minus Best fixed, and reports $\mathbb{E}[G]$ as an estimate because a sign-flip null is degenerate for a pathwise nonnegative quantity. Each contrast is the paired task difference averaged over all planned tasks, with early-terminal tasks contributing zero. We use a two-sided sign-flip test over floorplan-group sums with 50,000 Monte Carlo assignments and apply Holm correction within each family at familywise $\alpha = 0.05$. A positive estimate with an adjusted p value above 0.05 is not called an improvement.

The freeze is what makes the variance reduction safe. Fitting the machinery that lowers variance on the data it will later evaluate is enough to produce any desired conclusion: in extensive-form game evaluation, a heuristic value function tuned on the evaluation sample can be driven to pathologically low variance or to a chosen test statistic, so the function must be fixed before that sample is seen (Kim & Sandholm, 2026). Our controllers, transforms, and decision rules are hashed and committed before any confirmation outcome exists for the same reason. Panel size follows from the development operating point rather than from convenience. Treating each task as a paired binary comparison (Miller, 2024), the development contrast against continuation needs roughly a thousand analyzable tasks for conventional power, while the contrast against the best fixed repair needs an order of magnitude more; the second is therefore reported as an estimate and is not tested. Repeated continuations

are the variance reduction that makes the first contrast measurable at all, which is the same device Miller (2024) recommends for single-turn evaluations. Intervals are 10,000-resample percentiles at two levels: a task bootstrap and a grouped bootstrap over complete floorplans or ScienceWorld families. They are descriptive; the sign-flip and Holm procedure is the decision rule. The evaluator also reports every remaining contrast, intervention rate, action counts, recoveries, disruptions, and costs, and every ScienceWorld comparison is retained regardless of sign. No task is removed for a poor outcome, and missing eligible cells block final analysis. Appendix E gives the complete specification and Appendix D the administrative resumes.

5 RESULTS

5.1 SCHEDULED CHECKPOINTS AND THE FOUR-MESSAGE MENU

We estimate the direct effect against continuation at 1.56 percentage points (group-bootstrap 95% interval $[-0.12, 3.41]$; group sign-flip $p = 0.1048$; Holm-adjusted $p = 0.2096$). Against the matched arm-outcome comparator we estimate -0.39 points ($[-2.26, 1.57]$; adjusted $p = 0.7927$). The two controllers reach 54.43% and 54.82%. The denominator contains all 384 planned tasks: 355 reached the frozen checkpoint and 29 ended earlier, and Table 5 reports the full census. Neither primary controller comparison reaches significance at the unadjusted level, so neither is eliminated by the multiplicity correction alone. The development-selected safe policy changes success by 1.95 points relative to continuation. These tests answer the frozen comparisons; a favorable unadjusted subset is not substituted for them.

Direct selects continuation on 245/355 eligible prefixes, warning on 4, replanning on 106, and rollback on 0; it records 26 draw-level recoveries and 14 disruptions. Matched intervenes on 78.0% of prefixes, with 66 recoveries and 51 disruptions. The unconditional repair reaches 55.34% with 74 recoveries against 55 disruptions: Direct is the more precise rule and the unconditional rule is the more productive one, because firing on 31.0% of prefixes forgoes the net recoveries that firing everywhere collects. Direct uses 27.4 calls, 42560 input tokens, and 63.4 client-seconds per planned task, versus 28.1, 43557, and 64.9 for continuation. These summaries are descriptive. Section 6 reads the same branches for what this decision point and this menu could have delivered, which is what the second confirmation then changes.

5.2 EVENT-TRIGGERED CHECKPOINTS AND THE EXTENDED MENU

We find that cross-draw control is the strongest policy on the event-triggered panel, at 56.24% success against 54.97% for continuation and 54.30% for the best fixed repair. The denominator contains all 593 planned tasks: 497 reached a triggered checkpoint and 96 produced no trigger before termination. Against continuation we estimate 1.26 percentage points (floorplan-bootstrap 95% interval $[-0.35, 2.83]$; group sign-flip $p = 0.1579$; Holm-adjusted $p = 0.1579$), which the frozen test does not certify. Against the best fixed repair we estimate 1.94 points ($[-0.17, 4.17]$; $p = 0.0855$ (unadjusted)); we pre-specified that contrast as secondary and did not size the panel to resolve it.

The rule we deploy selects continuation on 298 of 497 triggered prefixes, the warning on 36, and the relocation message on 163; it leaves the remaining arms unused. It fires on 40.0% of triggered prefixes and records 52 draw-level recoveries against 37 disruptions, versus 105 and 113 for the unconditional repair. Tables 9 and 10 report every frozen policy and contrast, including the comparison-only and failure-risk learners that we run as the two closest external objectives inside this interface.

6 WHAT REPEATED BRANCHES REVEAL

We measure the same-draw action maximum against independent-draw evaluation. It exceeds it by 15.36 percentage points over all ALFWorld tasks (task-bootstrap 95% interval [12.50, 18.49]; floorplan-bootstrap [12.46, 18.30]; Holm-adjusted $p < 10^{-4}$). The gap is pathwise nonnegative. Its magnitude depends on how often stochastic branch outcomes disagree, so we do not read it as an intervention gain or as the regret of the learned controller. It is a function of that instability: the 179 tasks whose recorded cells never disagree contribute 0.00 points, the 95 tasks in the highest outcome-variance bin contribute 41.58 points, and the two correlate at 0.61. An environment whose

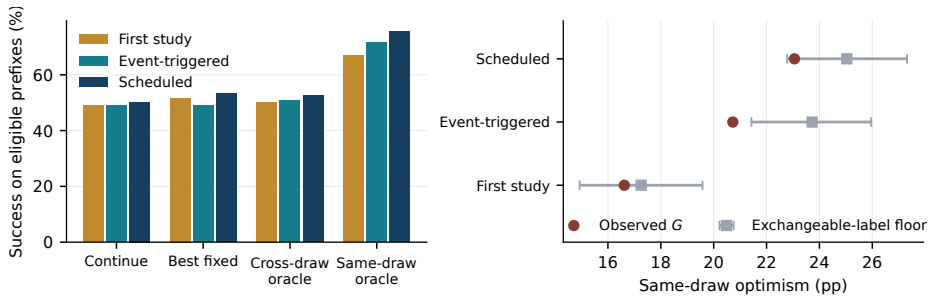

Figure 2: How much per-task signal the decision point and the action menu carry, and how much same-draw optimism is mechanical. Left: success on eligible prefixes for continuation, the best arm in hindsight, and per-task selectors scored across and within draws, for the first study's scheduled panel and both rules of the second. Right: observed same-draw optimism against the within-task exchangeable-label reference.

tools or simulated users add randomness therefore raises the diagnostic, so these estimates are a lower reference for such settings. We report the adjusted probability because the first study pre-registered it. A sign-flip null is degenerate for a pathwise nonnegative quantity, so we read the gap itself as an estimate against the reference distribution introduced below.

We connect the two confirmations with two readings, and we take both from the same repeated branches rather than from a new experiment. The first is how much per-task signal the menu carries. We give a per-task selector one complete independent draw of every arm and score it on the other draw. It reaches 50.80% on triggered prefixes against 49.09% for the best arm in hindsight, the warning, a margin of 1.71 points; under the scheduled rule the same quantities are 52.50% and 53.33% for the relocation message, a margin of -0.83 points. A trigger exists on 497/593 tasks, against 540/593 that reach the scheduled checkpoint on the same identities, so the event rule does not open its room by offering more chances to act. Relative to continuation the arm means are the warning 0.20, replanning -0.70, the one-step rollback -1.81, the three-step rollback -3.42, the relocation message -0.80 percentage points, so the menu is not interchangeable: the three-step rollback is the costliest arm on average and the warning the least costly. An average cost is not a verdict on any prefix, which is the distinction a controller is asked to make. Conditioning the decision on a public failure signal is what opens prefix-level room for a controller; a larger learner at the earlier decision point does not. The paired rules make that reading falsifiable, because the menu and the identities are the same on both. Where the reading is negative, average repair is what pays for a rule fitted at a different decision point: under the scheduled rule cross-draw control adds 0.51 points over continuation, while the failure-risk detector fires on 86.1% of prefixes and adds 3.71, which is close to the menu's unconditional effect rather than a prefix-level one. Where the instrument reports positive room the ordering reverses, and cross-draw control leads that detector by 1.85 points. The decision point, not the menu, decides what a controller has to condition on.

The second reading is how much of the same-draw gap is mechanical. We permute the recorded action and draw labels within each task, which fixes the continuation noise level and removes any action structure. Under the event rule the observed gap is 17.37 points over planned tasks and 20.72 points over eligible prefixes, against an exchangeable-label floor of 23.72 points ([21.43, 25.96]). The gap tracks continuation noise, which is why it is reported as an estimate rather than tested.

We can also move a controller to where the room is. We fitted the frozen rule on development records collected under the triggered rule and then read it under both, which is the decision-point mismatch that Section 2 warns about. We therefore refit the same learner on development records collected under the scheduled rule, and we break grid ties by how often the nested inner folds made the same choice rather than by the firing rate, which is the selection-frequency criterion of Meinshausen & Bühlmann (2010). That raises scheduled-rule success from 55.23% to 58.26% and puts cross-draw control ahead of every comparator on that panel as well: 3.54 points over continuation (adjusted $p = 0.0073$), 0.34 over the failure-risk detector, 0.17 over the repeated arm-outcome forest, and 0.76 over the best fixed repair. The rule it learns is the unconditional repair on 472 of 540 eligible prefixes together with a learned decline on 64: at a scheduled step the prefix says when not to repair rather

than which repair to use. The fit post-dates that panel's outcomes, so this reading is reported beside the frozen confirmation and not in place of it.

Two variants did not transfer, and we report both. Carrying the same static fallback to the triggered decision point moves success from 56.24% to 55.73%, which the interval of the contrast separating the two rules already contains. Gating the fallback on the development room reading inverts it. The development panels read -0.29 and 1.03 points where the confirmation panels read 1.71 and -0.83. At a few hundred tasks we therefore read the room on the panel we will deploy on rather than carry it from development.

7 Closest methods and comparison boundary

Same-prefix intervention learning. Calibration Is Not Control fits random-forest intervention policies from same-prefix branches, but evaluates its main controller with one realized outcome per action (Zhang et al., 2026). COTA trains a 0.5B pairwise advisor and reports 82.84–90.30% ALFWorld success with Qwen3-8B (Jiang et al., 2026b); its labels come from the same construction as ours, so the diagnostic of Section 2 applies to its supervision as well as to our evaluation, while its 30-step actor, environment-action candidates, repeated schedule, and offline budget differ from our one-shot menu, so its headline score is not a matched estimate of our contrast. Compared with both, this study values the option of not intervening on continuations that did not choose it. **Critic-based runtime supervision.** Asymmetric actor-critic supervision pairs a large actor with a small intervening critic (Jiang et al., 2026a), and AgentTether localizes a failure-critical subtrajectory and re-executes with scoped guidance (Zhao et al., 2026). Both choose what to say; neither estimates the utility of declining to intervene. Within our frozen interface we therefore run the comparison-only learner that COTA's objective implies and the failure-risk detector of Vasudev et al. (2026), matched on features, folds, capacity, actor revision, and decision budget.

Protecting a selected claim. The same discipline has been applied one level up. Selection-aware stress testing freezes a task reweighting chosen on discovery tasks and re-evaluates it on separate confirmation tasks, reporting no claim when support fails; on τ-bench with a Qwen3-8B actor a 3.75-point discovery gain vanished on confirmation (Xu et al., 2026). Compared with that protection, ours acts one level down: the unit of reuse is the continuation that scores an action rather than the task mixture that weights a benchmark, and the two compose.

Search-based repair. This is a different cost class. Language agent tree search expands and scores many continuations per decision and commits to the best (Zhou et al., 2024). That procedure is the deployable form of the outcome-informed selector in Equation 3, and Proposition 1 says exactly what it costs to score such a procedure on the branches that chose it. Our menu spends one suffix per task, so we report the selector as a diagnostic rather than as a baseline we outrank.

Estimation. Outcome modeling estimates every response function; direct signed-benefit modeling estimates effects relative to continuation (Künzel et al., 2019). Cross-draw scoring separates maximization from evaluation, following the principle behind Double Q-learning (van Hasselt, 2010), and Section 3.2 carries it into the controller. Inference-aware policy optimization characterizes the frontier between a learned policy's expected gain and the significance of that gain (Bastani et al., 2025); Section 6 reads our margin grid against that frontier and finds the frozen point at its most significant end. Every action here is executed from a serialized simulator state, so the estimand is the utility of the frozen menu, not a general causal effect or a benchmark rank. Our environments randomize only token sampling; Section 6 shows the diagnostic is a function of per-task outcome variance, so a domain whose tools and simulated users add randomness, as in τ-bench (Yao et al., 2024), moves it upward.

8 Limitations, ethics, and conclusion

Limitations. The confirmations cover one 8B actor, two text environments, six fixed repairs, one trigger rule, and two draws per action, and not web or code agents. The triggered gain over continuation is positive but not certified: the pre-registered calculation put the required panel near 1,100 identity-disjoint tasks and the balanced pool was exhausted at 593. The first confirmation's

null is bounded to a coarse menu at an early step, where room and capacity place it in the design. ScienceWorld is boundary evidence, reported in full and pooled with nothing.

The study uses simulated household and elementary-science tasks and contains no human-subject data. Interventions can still degrade task success, so the artifact reports disruptions and includes continuation as an action. Compute is capped and charged by GPU occupancy. The supplementary archive removes user names, host names, repository ownership, and absolute paths.

We introduced a repeated-branch protocol that separates the continuation used to choose a runtime repair from the one used to value it, and a reading of the same branches that says how much per-task signal a decision point and a menu carry before a controller is fitted. On the first frozen panel the protocol establishes a 15.36-point gap between same-draw maximization and independent-draw evaluation, and a within-task exchangeable-label reference places 17.26 points of it in continuation noise rather than in action advantage. Where a complete independent draw of the menu cannot beat the best arm in hindsight, the per-task outcome signal is thinner than the menu average, and on a panel run under both checkpoint rules with one menu that is what separates them. Moving the decision to the first public failure signal and separating repair depth from repair content lifts that reference to 50.80% against 49.09%, and a controller that chooses its action on one development draw and values it on another changes success by 1.26 points against continuation and by 1.94 points against the best fixed repair. Runtime controllers should include no action, freeze their choices before outcomes, value them on draws not used for selection, and read what the decision point and the menu carry before blaming the learner for a null.

Reproducibility statement. Appendix B gives the frozen prompts, Appendix C the controller grids, and Appendix H the task construction, seeds, software, container digest, and exact commands. The anonymous supplement includes the frozen configurations and predictions, raw result hashes, evaluator, figure builder, environment lock, and a script that regenerates every reported number, linked by SHA-256 manifests.

AI use statement. AI tools assisted literature search through alphaXiv, experiment and analysis code, drafting, and format checks, and procedural research skills shaped the review workflow (Kassis et al., 2026). Hypotheses, panels, inference, stopping rules, and controller predictions were frozen before their outcomes; quantitative claims were regenerated from stored artifacts and adverse results retained. The authors remain responsible for the manuscript.

REFERENCES

Hamsa Bastani, Osbert Bastani, and Bryce McLaughlin. Beating the winner's curse via inference-aware policy optimization. *arXiv preprint arXiv:2510.18161*, 2025.

Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On randomness in agentic evals, 2026. URL https://arxiv.org/abs/2602.07150.

Leo Breiman. Random forests. In *Machine Learning*, volume 45, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.02629.

Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Shuli Jiang, Zhaoyang Zhang, Yi Zhang, Shuo Yang, Wei Xia, and Stefano Soatto. Asymmetric actor-critic for multi-turn LLM agents, 2026a. URL https://arxiv.org/abs/2604.00304.

Yanze Jiang, Mingxuan Li, Yuhao Wang, Shengfang Zhai, and Jiaheng Zhang. Don't solve, just compare: Tiny advisors for runtime intervention in LLM agents, 2026b. URL https://arxiv.org/abs/2608.21027.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents, 2026. URL https://arxiv.org/abs/2609.00065.

Juho Kim and Tuomas Sandholm. Heuristic pathologies and further variance reduction via uncertainty propagation in the AIVAT family of techniques, 2026. URL https://arxiv.org/abs/2605.14261.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Hado van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Rakshith Vasudev, Melisa Russak, Dan Bikel, and Waseem Alshikh. Accurate failure prediction in agents does not imply effective failure prevention, 2026. URL https://arxiv.org/abs/2602.03338.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Yang Xu, Chenang Li, Jiefu Zhang, Haixiang Sun, Zhou Li, and Vaneet Aggarwal. Selection-aware stress testing for interactive agents, 2026. URL https://arxiv.org/abs/2608.30916.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Chubin Zhang, Zhenglin Wan, Xingrui Yu, Jingxuan Wu, Qi Wen, Pengfei Zhou, Wangbo Zhao, and Ivor Tsang. Calibration is not control: Why LLM-agent oversight needs intervention, 2026. URL https://arxiv.org/abs/2606.21399.

Chenyu Zhao, Shenglin Zhang, Wenwei Gu, Yongqian Sun, Dan Pei, Chetan Bansal, Saravan Rajmohan, and Minghua Ma. AgentTether: Graph-guided diagnosis and runtime intervention for reliable LLM agent operation, 2026. URL https://arxiv.org/abs/2607.06273.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.04406.

A Derivation of the repeated-branch diagnostic

For a fixed task, write $m^{(r)} = \max_a Y_a^{(r)}$ and $\widehat{a}^{(r)} = \arg\max_a Y_a^{(r)}$, breaking ties by the lowest action index. Because $a = 0$ carries the lowest index, continuation wins every tie in which it attains the maximum, and the lowest-indexed successful intervention wins otherwise. This convention fixes the realized value of G in the binary-outcome case, where ties are common. For either draw,

$$m^{(0)} - Y_{\widehat{a}^{(1)}}^{(0)} \geq 0, \qquad m^{(1)} - Y_{\widehat{a}^{(0)}}^{(1)} \geq 0, \tag{7}$$

because the selected action from the other draw remains an element of the maximization set. Their mean is G, so $G \geq 0$ pathwise. The result is unchanged by ties because any deterministic tie choice indexes an available action.

Now condition on the shared prefix $X = x$. Let the two action-outcome vectors be independent and identically distributed given $X = x$; independence is what allows the cross-draw term to factor, and identical distribution is what makes the two orderings interchangeable. Then

$$\mathbb{E}[G \mid x] = \mathbb{E}[m^{(0)} \mid x] - \mathbb{E}[Y_{\widehat{a}^{(1)}}^{(0)} \mid x] \tag{8}$$

$$= \mathbb{E}\Big[\max_a Y_a \mid x\Big] - \mathbb{E}\big[\mu_{\widehat{a}(Y')}(x) \mid x\big], \tag{9}$$

where Y' is an independent action-outcome vector and $\widehat{a}(Y')$ is the action selected from it. The first term is a realized maximum. The second is the value of the outcome-informed selection rule on a fresh continuation. It is not the value of any prefix-only deployable controller. If all actions share the same success probability but outcomes fluctuate, the difference may still be positive; G therefore identifies evaluation optimism rather than a causal intervention effect.

Two consequences follow for inference. First, a sign-flip permutation null is degenerate for G: non-negativity makes every group sum non-negative, the observed statistic is the algebraic maximum over sign vectors, and rejection only excludes deterministic continuations. We therefore report G as an estimate with bootstrap intervals and exclude it from the tested family. Second, the mechanical part of G can be quantified directly. Permuting the recorded outcomes within a task across all action and draw cells holds the number of successes fixed, preserves the continuation noise level, and destroys any action structure. The mean of G under this within-task permutation is the noise floor that a homogeneous menu would produce, and the observed G is read against it. With $R > 2$ draws we average G over ordered draw pairs, which reduces to Equation 3 at $R = 2$.

B Action and information contracts

Every arm receives the same task instruction, public action history, public observations, actor weights, decoding parameters, and total decision budget. An arm that reverts k decisions restores the serialized snapshot at step $t - k$ unconditionally, regardless of domain-level physical reversibility, and receives suffix limit $L - k$ for an episode limit L; the actor therefore makes exactly L decisions on every arm and $L - k$ environment actions remain in the episode. The controller receives only fields computed from the prefix. Environment state hashes, future observations, suffix outcomes, rollback metadata, and task-selection provenance are excluded. Arm 3 has no special access to a future legal-action sequence. The intervention decision occurs once per episode.

C Controller construction and development selection

The 50 numeric features are action-class fractions; counts and rates of admissible, repeated, invalid, and failed actions; reward summaries; observation and prompt lengths; token usage; action and observation novelty; vocabulary growth; and step index. Environment and checkpoint identity enter through the stratum: a separate forest is fitted for each actor and environment, and the checkpoint index is one of the 50 numeric fields. No categorical task identifier is encoded. The text field concatenates the recent public observations and actions. Word TF–IDF uses a vocabulary of at most 512 terms, and the resulting representation is reduced to at most 16 truncated-SVD components. All text fitting occurs inside each task-group fold during development selection. The final transform is refit on all admitted development tasks only.

Table 2: The complete frozen action menu. Arms 0–3 are the first confirmation's contract, unchanged; arms 4 and 5 are added by the second confirmation. The message is inserted once immediately before the next actor decision and remains in the visible transcript.

Arm	Name	Operation
0	CONTINUE	Restore the checkpoint and append no message.
1	WARNING	"Re-check the goal, the latest observation, and your previous action. Continue only after correcting any inconsistency you notice."
2	REPLAN	"Construct a fresh, concise plan from the current goal and observable state. Then execute that plan."
3	ROLLBACK–REPLAN	Restore the serialized environment and dialogue state at step $t-1$, append "The previous attempted action has been reverted. Construct a fresh, concise plan from the current goal and restored observable state. Then execute that plan," and set the suffix limit to $L-1$.
4	DEEP ROLLBACK–REPLAN	Restore the serialized environment and dialogue state at step $t-3$, clamped at the episode start, append "Your three most recent decisions have been reverted and the environment restored to the state before them. Construct a fresh, concise plan from the current goal and restored observable state. Then execute that plan," and set the suffix limit to $L-k$ for the k reverted decisions.
5	LOCATE	"Before acting again, name the objects this task still requires and the places you have not yet inspected. Then inspect the most promising place."

Table 3: Development-only policy selection for the first confirmation; the admitted development task count per environment is given in parentheses. Utilities average the two saved continuation draws within each complete task row. The strongest matched comparator and the operational selection were fixed before confirmation outcomes.

Environment	Candidate	OOF utility (%)	Intervention rate (%)
ALFWorld (129)	CONTINUE	45.35	0.00
	BEST FIXED (arm 2)	50.39	100.00
	DIRECT	47.67	34.11
	ARM OUTCOME (REPEATED)	47.29	96.90
	SINGLE-1	**51.16**	79.84
ScienceWorld (55)	CONTINUE	20.91	0.00
	BEST FIXED (arm 2)	24.55	100.00
	DIRECT	23.64	21.82
	ARM OUTCOME (REPEATED)	23.64	89.09
	SINGLE-1	**27.27**	100.00

The repeated arm-outcome control, REPEATED, predicts all four outcomes after fitting to the mean of two draws. SINGLE-1 uses the same action-outcome formulation but fits to one predesignated draw; it was the strongest runnable development comparator. DIRECT predicts three outcomes relative to continuation. Both target-parameterization learners share the same feature builder, task-group folds, 200-tree random-forest class, minimum-leaf grid $\{5, 10\}$, continuation-margin grid $\{0.00, 0.05, 0.10\}$, and task rows. A candidate must exceed the continuation margin to intervene; otherwise it chooses arm 0. CROSS-DRAW fits one signed-benefit forest per development draw on the same folds, features, and capacity, and adds two entries to the shared grid: an agreement requirement and a restriction of the candidates to the two arms with the highest mean benefit inside the training fold. With agreement required, it fires the common arm $\widehat{a}^{(0)} = \widehat{a}^{(1)}$ when the mean of the two cross-draw values exceeds the margin, and declines otherwise; this rule performs no maximization over scores that also value the chosen arm. Without agreement required, it keeps the better of the two cross-scored candidates, which is a maximization over two honestly cross-scored values rather than over the full menu. Both entries are scored by development out-of-fold utility inside the same frozen grid, and the candidate set is recomputed inside every training fold. We additionally report a nested out-of-fold utility that selects leaf, margin, and both added entries inside each training fold and scores the resulting rule on the held-out fold, which removes the optimism of selecting and scoring a grid on the same folds. Utility ties favor the lower firing rate and then the prespecified candidate order. The

released fitter reproduces this frozen selection by default; the static fallback and the nested-agreement tie-break are reached only through an explicit flag, and are reported as exploratory.

D PANEL CONSTRUCTION AND LINEAGE

The ALFWorld source inventory was grouped by its six task types. We removed every goal directory found in the 59 supplied prior configuration and outcome files, ranked the remaining identities by SHA-256 with seed 270914, and retained the first 64 identities per type. The resulting 384 tasks occupy 96 floorplan groups. ScienceWorld began from a 57-variation reserved official-test frame. We verified that none of its variations occurred among the 227 variations in the supplied exposure packet; the panel occupies 25 task families. These exclusions establish independence from the supplied research record, not from model pretraining or all work by other parties.

The second study draws from the same inventory after removing, in addition, every goal directory used by the first confirmation panel. Its confirmation panel holds 288 identities balanced across the six task types; its development panel holds 424 further identities and is disjoint from both confirmation panels. The two checkpoint rules run the same 288 confirmation identities, so the rule contrast is paired at the task level; because the frozen configuration hash differs between the rules, each rule generates its own baseline episode under the same task specification, seed, actor revision, and decoding parameters.

Table 4: Pre-outcome lineage. Full digests are machine-checked in the supplement.

Object	Frozen identifier
Panel and analysis commit	`8ddf66b` and `4dd0785` precede baseline generation
Prediction-freeze commit	`f0f26bf`, pushed before arm generation
Eligible prefixes	406: 355 ALFWorld and 51 ScienceWorld
Prefix digest	`0f5a80d1...46e863`
Prediction digest	`78380914...f6911b`
Arm-outcome family bundle digest	`759b9d02...7d612b`; it holds every arm-outcome learner, including REPEATED and the selected SINGLE-1
Direct model digest	`4ad68910...fe16db`

Baseline generation produced 35 structural early terminations. One episode exceeded its original coordinator lease. A single administrative resume reused the frozen task, seed, actor, software, and checkpoint rule; the interrupted attempt and resumed occupancy are both retained in the append-only ledger. No arm outcome was available when features and actions for the remaining 406 prefixes were serialized and committed. During arm generation, eight serving endpoints became unavailable after roughly one hour, leaving 560 canonical cells with technical failure flags. The evaluator rejected that state. The administrative resume archived every failed cell as an `attempt1` record and regenerated only those cells with their original configuration, prefix, action, draw, and request seeds. A separate server preflight failed on an occupied port and generated no outcome; restarting it on a different port completed the shard. No second-attempt record exists, and no failed cell enters the analysis.

E COMPLETE STATISTICAL SPECIFICATION

For a policy p, task-level utility is the mean of the two selected arm outcomes. Structural early-terminal tasks contribute their observed terminal success to every policy. A policy contrast is the paired difference in task utility. Task-bootstrap intervals resample tasks with replacement. Group-bootstrap intervals resample floorplans for ALFWorld and task families for ScienceWorld, retaining every task in each sampled group. Both use 10,000 draws and fixed stored seeds.

The primary randomization statistic is the absolute mean task contrast. Under the null, we independently multiply each group sum by -1 or $+1$, retaining within-group task dependence, and compare 50,000 Monte Carlo assignments with the observed statistic. The Monte Carlo p value includes the observed assignment through the usual plus-one correction. We order the raw ALFWorld p values within a family and apply Holm's step-down adjustment. The first confirmation's family comprises the optimism gap, DIRECT minus CONTINUE, and DIRECT minus SINGLE-1; the second confirmation's

family comprises CROSS-DRAW minus CONTINUE and CROSS-DRAW minus the development-best unconditional action, and collapses to the first of these when that action is continuation. The evaluator records the family it applied. All other tests and intervals are descriptive secondary analyses.

Recoveries count draw-level cases in which continuation fails and the policy succeeds. Disruptions count the reverse. Intervention rate is computed over eligible prefixes; success denominators include all planned tasks. Request, token, and client-wall-time summaries report the selected suffix for eligible tasks. The release additionally preserves all recorded calls and the GPU lease ledger. No missing eligible cell is imputed: the final evaluator refuses to emit results if any required arm-by-draw cell is absent.

F COMPLETE CONFIRMATION RESULTS

ScienceWorld provides a prespecified boundary rather than a second opportunity to claim success. Its continuation success rate is 19.30%, so most tasks fail under every action and contribute zero to both the gap and every contrast; the panel therefore compresses effects mechanically rather than showing greater branch stability. Its same-draw gap is 5.26 points ([1.69, 10.00] by task-family bootstrap), and Direct minus Continue is 0.00 points; its [0.00, 0.00] interval is degenerate because the controller fired on 11.8% of prefixes and produced no recovery or disruption, which is an absence of observed differences rather than an exact null. All 57 tasks, 25 families, six policies, costs, recoveries, and disruptions appear in Table 7 and the released result JSON. Differences in direction across environments narrow the empirical scope instead of motivating post hoc domain selection.

Table 5: Complete confirmation policy census. Success uses every planned task. Firing is over eligible prefixes; recoveries and disruptions count the two draw-level comparisons with continuation.

Domain	Policy	Success (%)	Fire (%)	Recover	Disrupt
ALFWorld	Continue	52.86	0.0	0	0
	Best fixed	55.34	100.0	74	55
	Direct	54.43	31.0	26	14
	Arm outcome	54.04	96.6	69	60
	Matched	54.82	78.0	66	51
	Safe selected	54.82	78.0	66	51
ScienceWorld	Continue	19.30	0.0	0	0
	Best fixed	19.30	100.0	5	5
	Direct	19.30	11.8	0	0
	Arm outcome	19.30	100.0	5	5
	Matched	14.04	100.0	2	8
	Safe selected	14.04	100.0	2	8

Figure 3 shows both marginal task and grouped uncertainty for the primary family and the selected deployment rule. The group intervals are wider whenever shared floorplans induce material composition sensitivity. Statistical significance is determined only by the frozen group sign-flip family, not by whether a plotted interval excludes zero.

Table 6: All ALFWorld contrasts. Values are percentage points. Holm adjustment applies only to the first two rows and the same-draw diagnostic in the text.

Contrast	Mean	Task 95%	Group 95%	Group p	Holm p
Direct − Continue	1.56	[0.00, 3.26]	[−0.12, 3.41]	0.1048	0.2096
Direct − Matched	−0.39	[−2.73, 1.95]	[−2.26, 1.57]	0.7927	0.7927
Direct − Arm outcome	0.39	[−2.21, 2.99]	[−1.82, 2.64]	0.8228	–
Direct − Best fixed	−0.91	[−3.39, 1.56]	[−3.09, 1.31]	0.4881	–
Safe selected − Continue	1.95	[−0.91, 4.82]	[−0.38, 4.26]	0.1311	–

ALFWorld same-draw optimism is 15.36 points with task-bootstrap interval [12.50, 18.49], group-bootstrap interval [12.46, 18.30], raw group sign-flip $p =< 10^{-4}$, and Holm-adjusted $p =< 10^{-4}$. ScienceWorld optimism is 5.26 points with task-family-bootstrap interval [1.69, 10.00] and unadjusted group sign-flip $p = 0.0632$.

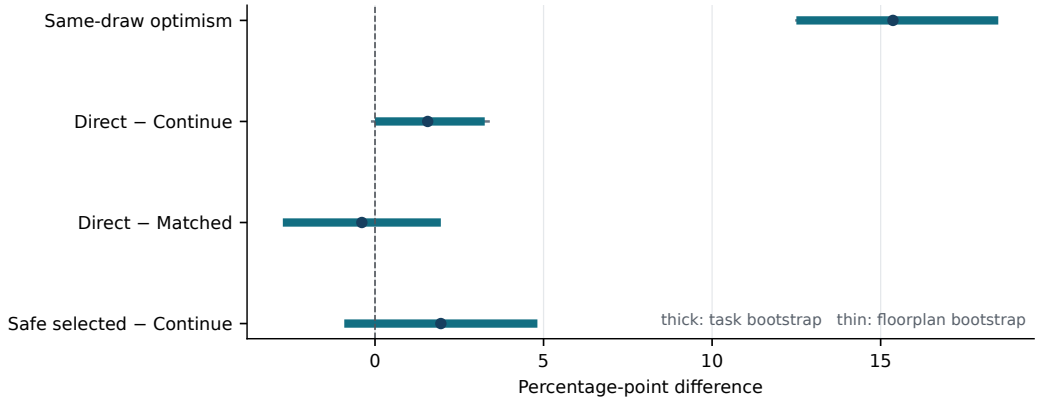

Figure 3: Confirmation effects on ALFWorld. Dots show planned-task means; thick and thin intervals show task- and floorplan-bootstrap 95% intervals. Reported adjusted p values use the prespecified group sign-flip tests and Holm correction.

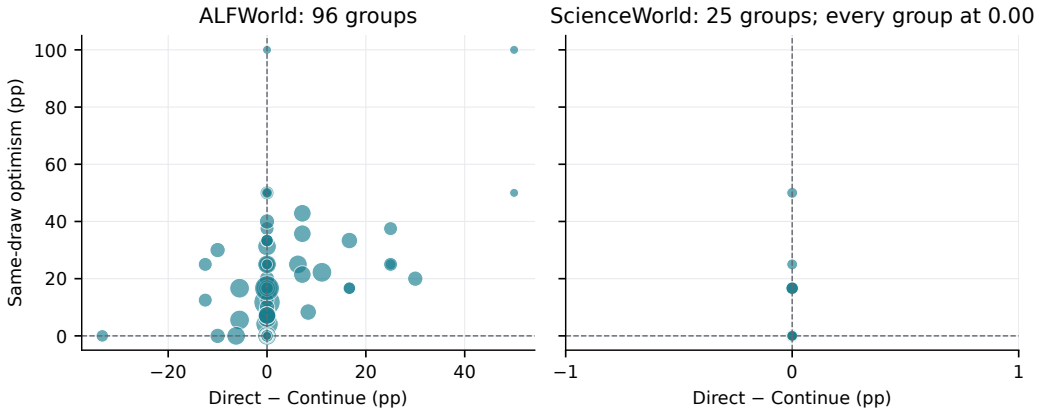

Figure 4: Group-level measurement and policy effects. Each point is a complete floorplan or task-family group; area is proportional to its task count. Vertical position is same-draw optimism and horizontal position is Direct minus Continue; a group in the upper left shows optimism from the outcome-informed selector while the deployable controller loses utility. Dashed lines mark zero.

Across baseline, arm, and retained attempt records, the episode ledger contains 127,213 model requests, 253,758,979 input tokens, 6,921,686 output tokens, and 301,110.6 client-seconds. It records 561 failed requests and 561 requests with unknown token usage. The separate GPU lease ledger charges model loading, qualification probes, generation, idle occupancy, failures, and cleanup.

We also raised learner capacity directly, on development data and before any test prediction. Nested honest out-of-fold utility is 29.47% for the frozen forest at the scheduled point and 50.69% at the triggered one; gradient boosting, appended 0.6B language-model embeddings of the prefix, and their combination each score lower at both (29.27, 29.10, 28.84 and 49.46, 50.29, 50.15). What the prefix carries at a decision point, not the model class, decides what a learner recovers. The same grid traces the trade-off between expected gain and the significance of that gain. Across the frozen candidate set the development improvement over continuation reaches 3.09 points and the paired z statistic reaches 2.60; the configuration the pre-registered utility criterion selected is the one attaining that maximum, at 2.94 points over 11.5% discordant tasks. The frozen operating point is the most certifiable one the menu offers, not the most aggressive.

Table 7: All ScienceWorld contrasts, retained as unadjusted boundary analyses. Values are percentage points. Identical rows are not typesetting errors: on this panel the arm-outcome forest selected arm 2 on every eligible prefix and therefore coincides with BEST FIXED, and development selection chose SINGLE-1, so SAFE SELECTED coincides with MATCHED.

Contrast	Mean	Task 95%	Group 95%	Group p	Holm p
Direct − Continue	0.00	[0.00, 0.00]	[0.00, 0.00]	1.0000	–
Direct − Matched	5.26	[−0.88, 12.28]	[−0.81, 11.86]	0.1860	–
Direct − Arm outcome	0.00	[−5.26, 5.26]	[−4.17, 4.39]	1.0000	–
Direct − Best fixed	0.00	[−5.26, 5.26]	[−4.17, 4.39]	1.0000	–
Safe selected − Continue	−5.26	[−12.28, 0.88]	[−11.86, 0.81]	0.1860	–

Table 8: Mean selected-suffix cost per planned task. Structural early terminations have zero suffix cost. Policies that select the same action on every eligible prefix share a row by construction.

Domain	Policy	Requests	Input tokens	Output tokens	Client seconds
ALFWorld	Continue	28.1	43557	1508	64.9
	Best fixed	27.5	43237	1468	63.2
	Direct	27.4	42560	1470	63.4
	Arm outcome	27.7	43559	1481	63.8
	Matched	27.7	43459	1487	64.1
	Safe selected	27.7	43459	1487	64.1
ScienceWorld	Continue	47.7	177369	2836	124.6
	Best fixed	48.3	193820	2837	126.5
	Direct	47.7	178377	2847	125.3
	Arm outcome	48.3	193820	2837	126.5
	Matched	52.8	199333	3139	138.0
	Safe selected	52.8	199333	3139	138.0

G HISTORICAL AND EXPLORATORY EVIDENCE

Earlier work motivated the new confirmation but does not enter its multiplicity family. On a 134-game official unseen ALFWorld panel with a different actor interface, the prior repeated-branch learner reached 68.16% against 69.40% for continuation, a difference of -1.24 percentage points with task-bootstrap interval $[-4.98, 2.24]$ and $p = 0.5946$. Same-arm mismatches ranged from 8.25% to 12.06%, and the same-draw sample oracle exceeded a fresh-replica evaluation by 7.94 points. A four-stratum retrospective direct-benefit reanalysis averaged 30.14%, compared with 32.62% for continuation and 28.92% for the repeated-outcome policy. These records motivated independent continuation evaluation and the direct target comparison; they do not confirm either new hypothesis.

The native-interface follow-up and its demonstration ablation are reported only as actor-side exploratory analyses. Changing thinking mode, output cap, and demonstration use together produced 80.60% versus 70.90% in one paired 134-task draw. A later conditional ablation held thinking and cap fixed: including two released demonstrations produced 116/134 successes (86.57%) versus 107/134 (79.85%), a $+6.72$ percentage-point estimate with task-bootstrap interval $[-0.75, 14.18]$ and two-sided sign-flip $p = 0.1087$. Excluding one scene reversed the latter estimate to -3.51 points. Neither result is pooled with confirmation, establishes a controller effect, or supports a state-of-the-art claim.

A separate 27B experiment is retained as a systems result. On its internal reserved panel, the plain 27B actor scored 131/160 (81.88%) against 76/160 (47.50%) for an 8B self-check system; the paired difference was $+34.38$ percentage points with a prespecified two-sided 97.5% cluster-t interval $[24.20, 44.55]$, whose lower limit is the one-sided 98.75% bound recorded for that systems check; every other interval in this paper is reported at 95%. The systems differ in model, policy, filtering, and decoder stream, and 49 of 64 confirmation floorplans had appeared in historical tasks. On the historically exposed 134-task public ALFWorld frame, the plain and self-check 27B systems both scored 108/134. These results establish neither an isolated parameter-count effect nor a controller advance and are not pooled with confirmation. A separate internal retrieval study on an unrelated corpus is outside this paper.

Table 9: Complete policy census under both checkpoint rules. Success uses every planned task; firing is over eligible prefixes; recoveries and disruptions count draw-level comparisons with continuation.

Rule	Policy	Success (%)	Fire (%)	Recover	Disrupt
Event-triggered	Continue	54.97	0.0	0	0
	Best fixed	54.30	100.0	105	113
	Direct	56.24	54.1	62	47
	Arm outcome	54.81	10.1	10	12
	Matched	55.06	0.8	3	2
	Comparison-only	55.48	54.9	60	54
	Failure risk	54.38	81.9	94	101
	Lower bound	54.97	0.0	0	0
	Cross-draw	56.24	40.0	52	37
	Safe selected	56.24	40.0	52	37
Scheduled	Continue	54.72	0.0	0	0
	Best fixed	57.50	100.0	140	107
	Direct	55.65	35.7	52	41
	Arm outcome	54.72	2.8	2	2
	Matched	54.72	0.4	0	0
	Comparison-only	55.99	37.8	49	34
	Failure risk	58.43	86.1	130	86
	Lower bound	54.72	0.0	0	0
	Cross-draw	55.23	26.3	37	31
	Safe selected	55.23	26.3	37	31

Table 10: Complete contrast census on the event-triggered panel. Intervals are 95% task and floorplan bootstraps; adjusted values apply Holm correction within the pre-registered family.

Contrast	Estimate (pp)	Task 95%	Group 95%	Raw p	Adjusted p
Cross-draw − Continue	1.26	[−0.51, 3.04]	[−0.35, 2.83]	0.1579	0.1579
Cross-draw − Best fixed	1.94	[−0.17, 4.05]	[−0.17, 4.17]	0.0855	–
Cross-draw − Direct	0.00	[−1.43, 1.43]	[−1.48, 1.57]	1.0000	–
Cross-draw − Arm outcome	1.43	[−0.17, 3.04]	[−0.09, 2.94]	0.0957	–
Cross-draw − Matched	1.18	[−0.51, 2.87]	[−0.35, 2.68]	0.1671	–
Cross-draw − Comparison-only	0.76	[−0.93, 2.36]	[−0.82, 2.43]	0.4226	–
Cross-draw − Failure risk	1.85	[−0.25, 3.88]	[−0.24, 4.05]	0.1001	–
Cross-draw − Lower bound	1.26	[−0.51, 3.04]	[−0.35, 2.83]	0.1579	–
Direct − Continue	1.26	[−0.59, 3.20]	[−0.79, 3.15]	0.2630	–
Comparison-only − Continue	0.51	[−1.43, 2.45]	[−1.60, 2.50]	0.6892	–
Failure risk − Continue	−0.59	[−2.95, 1.77]	[−3.45, 2.11]	0.7244	–
Lower bound − Continue	0.00	[0.00, 0.00]	[0.00, 0.00]	1.0000	–
Safe selected − Continue	1.26	[−0.51, 3.04]	[−0.35, 2.83]	0.1579	–

A temperature ablation runs the same checkpoint rule, menu, and draw count twice on 91 shared identity-disjoint ALFWorld tasks, once at the frozen sampling temperature of 0.7 and once at 1.0. It separates the diagnostic from the knob usually assumed to drive it. At 0.7 the mean per-task outcome variance is 0.0936 and the same-draw gap is 21.98 percentage points; at 1.0 the variance is 0.0796 and the gap is 16.48 points. Raising the temperature did not raise the variance on these tasks, and the gap moves with it. Temperature is therefore an indirect and non-monotone handle: what the diagnostic follows is the variance of the continuation outcome itself, which the profile in the main text measures directly within a single panel. A domain whose tools or simulated users inject randomness of their own raises that variance and so raises the gap, which is the sense in which the estimates here are a lower reference for such settings. This ablation is exploratory: its panel is small, it was generated after the confirmation panels, and it enters no confirmatory family.

The variance profile reported in Section 6 and this ablation answer the same question from two directions: within one panel the diagnostic rises with per-task outcome variance, and across a matched pair of panels it moves with the sampling temperature that produces that variance. The

Table 11: Pre-outcome lineage for the event-triggered study. Both checkpoint rules run the same 288 identities; predictions were frozen after baselines and before any arm outcome.

Object	Frozen identifier
Event-triggered eligible prefixes	497
Event-triggered prefix digest	`d20c439a...ba0775`
Event-triggered prediction digest	`738b1ff3...8ec801`
Scheduled eligible prefixes	540
Scheduled prefix digest	`02fc748b...332f88`
Scheduled prediction digest	`a869e454...1e3d43`

released evidence index preserves those boundaries rather than using favorable actor or domain subsets to replace the frozen confirmation.

H REPRODUCTION

The anonymous supplement contains the exact ICLR source, panel manifests, frozen prefix features and actions, controller manifests, raw confirmation records, analysis environment lock, tests, evaluator, rendering program, and artifact manifest. Model weights and benchmark installations are excluded. Analysis and paper regeneration require no GPU.

```
uv venv --python 3.12 .venv
uv sync --frozen
.venv/bin/python -m pytest -q
mv artifacts/confirmation artifacts/confirmation.release
.venv/bin/python controller/evaluate_confirmation.py \
  --config configs/independent-panel/panel-shard0.json \
  --config configs/independent-panel/panel-shard1.json \
  --config configs/independent-panel/panel-shard2.json \
  --config configs/independent-panel/panel-shard3.json \
  --config configs/independent-panel/panel-shard4.json \
  --config configs/independent-panel/panel-shard5.json \
  --config configs/independent-panel/panel-shard6.json \
  --config configs/independent-panel/panel-shard7.json \
  --raw raw \
  --predictions artifacts/prediction-freeze/predictions.json \
  --prediction-freeze \
    artifacts/prediction-freeze/prediction-freeze.json \
  --out artifacts/confirmation
.venv/bin/python paper/build_assets.py
cd paper
mkdir -p build
pdflatex -interaction=nonstopmode -halt-on-error \
  -output-directory=build main.tex
BIBINPUTS=.: BSTINPUTS=.: bibtex build/main
pdflatex -interaction=nonstopmode -halt-on-error \
  -output-directory=build main.tex
pdflatex -interaction=nonstopmode -halt-on-error \
  -output-directory=build main.tex
cd ..
.venv/bin/python tools/verify_artifacts.py --supplement \
  ../intervene-or-continue-iclr2027-supplement.zip
```

The second study reuses the same programs. Its panels are frozen with `extension/freeze_panel.py` under `-checkpoint-rule event` and `-checkpoint-rule scheduled` on one shared set of identities, its development rows are built by `extension/analysis.py`, its three learners are fitted by `extension/policies.py`,

`controller/fit_direct.py`, and `controller/fit_cross_draw.py`, its predictions are frozen by `controller/freeze_predictions.py` with the cross-draw bundle attached, and its results are produced by `controller/evaluate_confirmation.py` with `-primary CROSS_DRAW_vs_CONTINUE -primary CROSS_DRAW_vs_BEST_FIXED`. Tables and figures for both studies regenerate through `paper/build_assets.py` and `paper/build_assets_r2.py`.

Inference reproduction additionally requires ALFWorld 0.4.2, TextWorld 1.7.0, ScienceWorld 1.3.0, the recorded Qwen3-8B revision, and the content-addressed serving image. Each GPU worker receives one panel shard and a unique port. A bounded coordinator enforces a 72,000 H100 GPU-second actual-or-reserved occupancy cap; closed leases return unused reservation headroom. The immutable ledger records 47,876.7 charged H100 GPU-seconds across 26 closed leases. Exact request seeds are derived from task identity, arm, draw, and request index. New stochastic generations are replications and must not overwrite the released outcomes.

The supplement's verification program checks the source commit, all protected input hashes, complete task and cell counts, task/fold disjointness, action visibility, panel exclusions, prediction freeze, recomputed result JSON, figure inputs, page limit, unresolved references, anonymous strings, and final archive digest. Absolute execution paths and private repository ownership are replaced by symbolic roots in the anonymous archive; scientific identifiers and file-content digests are unchanged.

I AI ASSISTANCE AND AUTHOR RESPONSIBILITIES

AI tools were used for literature discovery through alphaXiv, code drafting, statistical and manuscript review, LaTeX drafting, copyediting, and format inspection. Manuscript structure was checked against five pre-2024 ICLR papers on grounded language agents and prompting, read from their arXiv sources: BabyAI, ALFWorld, self-consistency, least-to-most prompting, and ReAct. The released audit records their abstract and section lengths, their float conventions, a per-section profile of sentence length, first-person density and hedging, and a vocabulary comparison in which the reference papers cover one another at 79 to 89 percent of tokens while this manuscript reaches 71 percent. The shortfall is carried by the study's own objects, which are retained; no word was substituted to move that statistic, and no perplexity target or author imitation was used. They were not treated as authors. The pre-outcome commits and immutable result evaluator separate assistance during design from outcome-conditioned reporting. The complete result census, adverse evidence, source links, and assistance ledger are provided for audit. Human authors remain responsible for checking claims, citations, authorship, conflicts, affiliations, eligibility, disclosure categories, and the final submission.