DISAGREEMENT IS NOT GUIDANCE: AUDITING AND SELECTING SUPERVISION IN AGENTIC ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) uses teacher supervision on states visited by the student. Action- distribution disagreement alone gives little information about whether teacher guidance improves environmental utility. We propose Behavioral Necessity Audit (BNA), which evaluates student and teacher distributions on a fixed executable candidate set and shared continuation policy. BNA jointly measures action-distribution distance, executable outcome distance, and the conditional utility gain from shifting probability toward the teacher. On frozen ALFWorld and official full WebShop populations, outcome-aware transport costs are 93.16% and 93.40% lower than identity-based costs, while beneficial guidance occurs on 17.44% and 35.07% of states. We then propose Horizon-Split Necessity Gating (HSNG), which uses a short executable prefix to select teacher targets. H16 identifies terminally beneficial guidance with AUROCs of 0.8595 and 0.8693 on the original populations. On disjoint follow-up populations of 80 new states from 40 new task clusters per environment, the frozen rule reaches AUROCs of 0.80 and 0.79, with positive accepted terminal gains of $+0.10$ and $+0.07$. In a matched, single-seed ALFWorld micro-update, HSNG reduces mean off-selection KL drift by 90.60% while all six selected states move toward their teacher targets. These results ground supervision allocation in conditional environmental utility rather than surface action disagreement.

1 INTRODUCTION

Interactive agents driven by large language models (LLMs) typically rely on a stronger teacher model for knowledge transfer. On-policy distillation (OPD) applies teacher supervision to the student's own online trajectories, which can reduce distribution shift between offline demonstrations and the student's online behavior. It is a common paradigm for transferring the decision-making capabilities of stronger models to lightweight agents (Agarwal et al., 2024; Lauffer et al., 2026). However, OPD and the selective guidance methods that follow it often rely on an implicit assumption that has rarely been tested directly: when the teacher and student disagree on an action, introducing teacher supervision will usually be beneficial. Existing methods allocate supervision using turn-level disagreement or uncertainty (Tan et al., 2026), curriculum rules (Li et al., 2026), hindsight error localization (Yeo et al., 2026), or privileged-state matching (Liu et al., 2026). These mechanisms identify forms of student–teacher mismatch, while leaving the utility of shifting the student policy toward the teacher unmeasured in the interactive environment.

In interactive environments, different action paths can reach similar task outcomes, while a teacher-preferred action can also lead to a continuation that is worse than the student's. Action-distribution disagreement captures surface variation, whereas guidance value depends on the downstream consequences of the action distribution. Determining whether a state warrants teacher supervision therefore requires a counterfactual evaluation: how much does expected environmental utility change when the student's action distribution is replaced by the teacher's?

To answer this question, we introduce BEHAVIORAL NECESSITY AUDIT, a distribution-level protocol for evaluating behavioral necessity against environmental outcomes. BNA uses a frozen executable candidate set and a shared continuation policy to jointly measure student–teacher action-distribution

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.

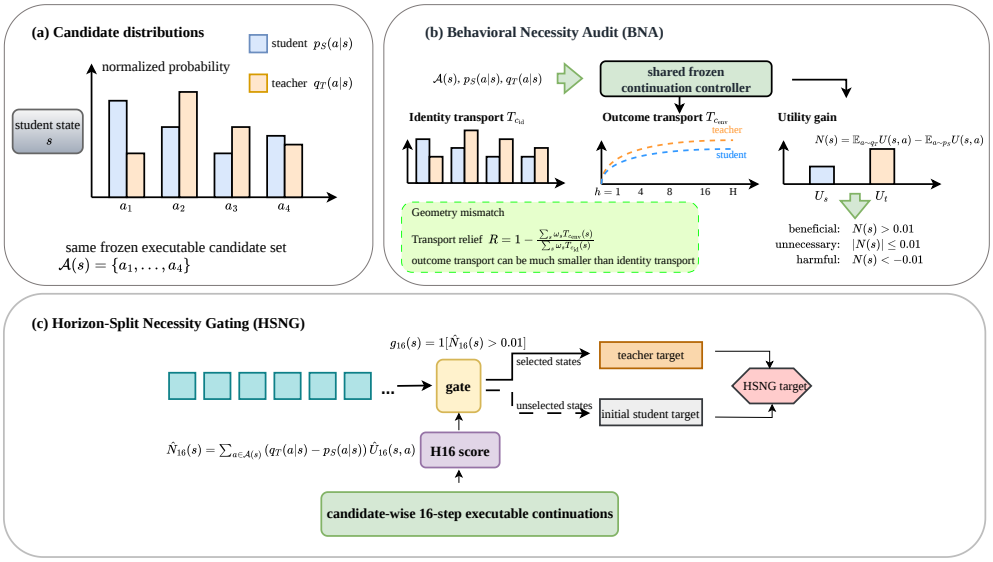


Figure 1: **Overview of the BNA and HSNG pipeline.** (a) Student and teacher distributions are normalized on the same frozen executable candidate set. (b) BNA measures identity transport, outcome transport, and directional utility gain, then assigns a behavioral necessity label. (c) HSNG uses the 16-step executable-prefix score to select teacher targets and anchors unselected states to the initial student distribution. Terminal BNA labels are used to evaluate H16 and are not inputs to the gate.

disagreement, executable outcome distance, and the conditional utility gain from the teacher relative to the student. This separates surface action differences from supervision that improves environmental utility. On frozen ALFWorld (Shridhar et al., 2021) and the full official WebShop benchmark (Yao et al., 2022), outcome-aware transport costs are 93.16% and 93.40% lower than identity-based transport costs, respectively. Yet beneficial guidance accounts for only 17.44% and 35.07% of states under state-pair weighting, indicating that imitating the teacher does not provide a clear benefit in most states and can even be harmful in some.

Full counterfactual rollouts are expensive and difficult to use directly for supervision selection during training. We therefore introduce H16, a score computed from a 16-step executable prefix that estimates guidance benefit at the terminal outcome. H16 achieves AUROCs of 0.8595 and 0.8693 on the original evaluation populations, and 0.80 and 0.79 on independent new tasks. Based on this identifier, we propose Horizon-Split Necessity Gating (HSNG), which applies teacher targets only to states with high predicted benefit and retains the initial student policy as the target for all other states. In a matched parameter-update experiment, HSNG reduces mean policy KL drift on unselected states by 90.60%, while still moving selected states toward the teacher policy.

Figure 1 illustrates the full pipeline.

Our contributions are:

- **An outcome-grounded behavioral necessity audit.** We introduce a distribution-level framework that disentangles action disagreement, executable outcome distance, and conditional utility gain, and show a clear mismatch between action differences and actual supervision value on ALFWorld and WebShop.

- **An outcome-driven selective supervision approach.** We show that a short executable prefix can identify states where supervision is highly beneficial, and use HSNG to study how supervision placement relates to policy alignment on selected states and policy drift on unselected states.

2 BEHAVIORAL NECESSITY

2.1 THREE DISTINCT OBJECTS

Let $p_S(a \mid s)$ and $q_T(a \mid s)$ be student and teacher distributions over a frozen executable candidate set $\mathcal{A}(s)$. We obtain sequence probabilities by scoring the complete candidate action, then normalizing within $\mathcal{A}(s)$. The usual exact-action geometry uses

$$c_{\mathrm{id}}(a_i, a_j) = \mathbf{1}[a_i \neq a_j]. \tag{1}$$

Given a ground cost c, we first obtain an entropically regularized coupling (Cuturi, 2013),

$$\begin{aligned} \pi_c^{\varepsilon}(s) &\in \arg\min_{\pi \in \Pi(p_S, q_T)} \left[\sum_{i,j} \pi_{ij} c(a_i, a_j) + \varepsilon \sum_{i,j} \pi_{ij} (\log \pi_{ij} - 1) \right], \\ T_c(s) &= \sum_{i,j} \pi_{c,ij}^{\varepsilon}(s) c(a_i, a_j). \end{aligned} \tag{2}$$

where $\varepsilon = 0.05$. We report the coupling-weighted ground cost T_c. Identity cost captures whether action strings match; c_{env} captures whether their environmental consequences match.

For each candidate, we execute paired continuations and record outcome curves over horizons $h \in \{1, 4, 8, 16, H\}$: validity, success, progress, time-to-success, terminal failure, and censoring. The environment cost c_{env} is the equally weighted mean of normalized validity, success-curve, progress-curve, and time/failure distances. We summarize how much surface geometry overstates consequence geometry by

$$\mathrm{TransportRelief}(\mathcal{S}; \omega) = 1 - \frac{\sum_{s \in \mathcal{S}} \omega_s T_{c_{\mathrm{env}}}(s)}{\sum_{s \in \mathcal{S}} \omega_s T_{c_{\mathrm{id}}}(s)}. \tag{3}$$

The aggregates are formed before taking the ratio. The resulting quantity is a relative distributional cost reduction.

Finally, behavioral similarity is symmetric, whereas guidance benefit is directional. With our frozen terminal utility

$$U(s, a) = \tfrac{1}{2} P(\mathrm{success} \mid s, a) + \tfrac{1}{2} \mathrm{progress}(s, a), \tag{4}$$

we define

$$N(s) = \mathbb{E}_{a \sim q_T} U(s, a) - \mathbb{E}_{a \sim p_S} U(s, a). \tag{5}$$

For the preregistered margin $m = 0.01$, guidance is *beneficial* if $N(s) > m$, *harmful* if $N(s) < -m$, and *unnecessary* otherwise. "Harmful" therefore means that the teacher *distribution* has lower estimated expected utility under the frozen evaluator. Every label is conditional on the candidate support, continuation policy, horizon and utility. Values inside the margin remain unclassified; the margin is a decision threshold rather than an equivalence test.

2.2 WHY DISAGREEMENT PRESSURE CAN BE MISALLOCATED

An exact OPD update can be large whenever p_S and q_T disagree, including states with small conditional guidance benefit. We call the share of a nonnegative state-level selector or loss pressure $w(s)$ allocated to non-beneficial states

$$\mathrm{NBP}(w) = \frac{\sum_s \omega_s w(s) \mathbf{1}[N(s) \leq m]}{\sum_s \omega_s w(s)}, \tag{6}$$

where ω_s is either the state-selection weight or decision-occupancy weight. NBP quantifies how selector pressure is distributed across beneficial and non-beneficial states.

3 AUDIT PROTOCOL

Outcome-independent populations. ALFWorld uses 120 states in 71 task clusters, selected by a frozen stratified hash from 311 unique states, spanning task family, depth, teacher gap, and action-set size. WebShop uses 100 states in 73 task clusters, selected from a frozen 162-state census in the official full product environment. Selection never reads candidate outcomes or necessity labels. We report a state estimand and an on-policy decision-occupancy estimand separately.

Table 1: **Behavioral necessity across formal outcome populations.** Shares and mean gain are state-weighted; brackets are 95% task-cluster bootstrap intervals. Transport relief is a distribution-level geometry statistic (Eq. 3).

Metric	ALFWorld	WebShop
States / task clusters	120 / 71	100 / 73
Transport relief	93.16% [90.10, 95.98]	93.40% [91.99, 94.70]
Beneficial	17.44% [10.94, 24.28]	35.07% [25.01, 45.30]
Unnecessary	65.29% [53.86, 76.48]	23.40% [14.91, 32.70]
Harmful	17.27% [10.00, 25.12]	41.53% [31.23, 52.42]
Mean teacher gain	-0.0235 [-0.0553, .0038]	.0027 [-0.0326, .0383]
SOFT-OR AUROC	.406 [.259, .558]	.385 [.273, .502]
Non-beneficial pressure	81.49% [73.03, 89.60]	68.09% [45.19, 85.91]

Candidate support. For ALFWorld, we score every official executable action (10–58 per state) and adaptively take the union of teacher and student top-K supports. All states stop by $K = 4$ with 4–8 retained actions and at least 0.80 captured mass for each policy. For WebShop, the candidate set contains all executable click actions plus frozen student/teacher search proposals; the same 0.80 support-mass rule is enforced, with coverage exclusions reported explicitly.

Counterfactual continuations. Each retained candidate is restored from the same snapshot and continued by a frozen Qwen3-32B observable-only controller. The prompt retains the task goal, recent action–observation history, action-derived inventory/location, and loop-aware replanning; no privileged simulator state enters the policy input. ALFWorld executes one greedy and three paired sampled continuations; WebShop uses the frozen full-horizon continuation protocol. The same logical seeds are paired across candidates. One trace supplies all horizons, which prevents horizon-specific resampling.

Models and integrity. The student is Qwen3-1.7B and the teacher is Qwen3-32B, both with thinking disabled and full action-sequence likelihoods. ALFWorld contains 120 audited primary records and WebShop contains 100; sample selection is frozen before outcome labels are joined.

Statistics. We use task cluster as the resampling unit and report 10,000 bootstrap replicates. State and decision weights flow through means, shares, pressure, and AUROC; each interval follows the estimator and population used for its statistic.

4 GEOMETRY MISMATCH AND GUIDANCE NECESSITY

4.1 SURFACE DISAGREEMENT OVERSTATES BEHAVIORAL SEPARATION

Figure 2a and Table 1 show the central result. ALFWorld has state-weighted identity and environment transport costs of 0.9196 and 0.0629; WebShop has 0.9984 and 0.0659. The corresponding transport relief is 93.16% and 93.40%. Thus, under the frozen counterfactual evaluator, identity transport substantially exceeds executable outcome transport. Optimal transport couples probability across several behaviorally nearby candidates, so the statistic describes distribution-level geometry.

To test whether this result is driven entirely by states whose terminal utilities are constant across candidates, we post hoc retain states whose within-state utility range exceeds 10^{-12}. This gives 55 ALFWorld states in 36 task clusters and 93 WebShop states in 68 task clusters. Using the same weighting definition, transport relief remains 86.21% and 93.05%, respectively. The contraction persists where candidate terminal utilities vary. Eq. 5 then determines whether guidance points toward or away from the teacher.

4.2 GUIDANCE BENEFIT VARIES ACROSS STATES

The directional result complements the geometry. Beneficial guidance appears in 17.44% of ALFWorld states and 35.07% of WebShop states. In ALFWorld, unnecessary guidance dominates

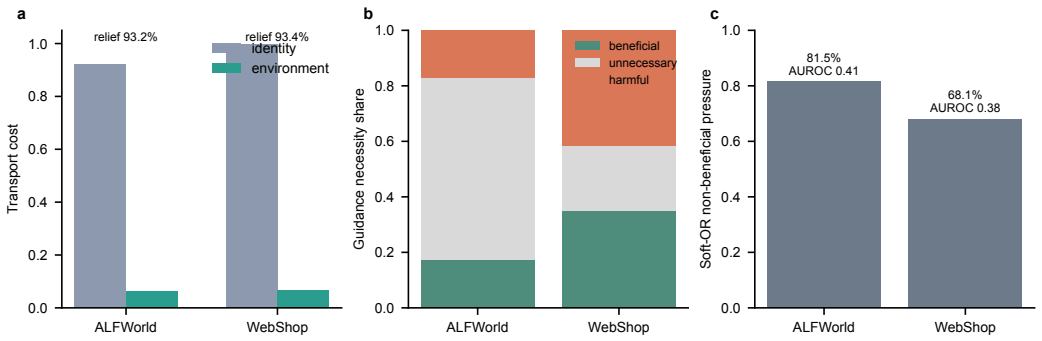


Figure 2: **Disagreement is not guidance.** (a) Identity transport is nearly maximal, while outcome-aware transport is small in both environments. (b) Guidance necessity under the frozen utility: beneficial states are a minority; unnecessary and harmful are distinct. (c) A likelihood SOFT-OR proxy allocates most pressure to non-beneficial states. AUROC is for identifying beneficial guidance. Values are state-weighted point estimates.

(65.29%); in WebShop, harmful guidance is the largest class (41.53%). Mean teacher-minus-student utility is close to zero at the population level. Top-action comparisons provide a consistent auxiliary view: the teacher top action is better on 12.40% of ALFWorld states and 35.99% of WebShop states, while the student top action is better on 14.85% and 39.32%, respectively. Distributional necessity also accounts for probability mass beyond the top action.

For an additional cross-check, WebShop contains 35 beneficial, 24 unnecessary, and 41 harmful states by unweighted count. Under decision-occupancy weighting, the mean teacher gain is 0.002377 with a 95% interval of $[-0.033915, 0.039045]$, showing substantial state heterogeneity.

4.3 LIKELIHOOD PRESSURE MISSES THE BENEFICIAL SUBSET

We evaluate disagreement, student entropy, and their continuous SOFT-OR combination as outcome-blind proxies. On WebShop their state AUROCs are 0.389, 0.469, and 0.385; all 95% intervals include 0.5. ALFWorld SOFT-OR AUROC is 0.406. SOFT-OR allocates 81.49% of ALFWorld pressure and 68.09% of WebShop pressure to unnecessary or harmful states (Figure 2c). The corresponding non-beneficial state shares are 82.56% and 64.93%. Together, the identification intervals and allocation statistics show that these outcome-blind proxies rank beneficial states unreliably on the audited populations.

Additional signal and construct analyses. Appendix A audits bounded reconstructions of concurrent selector primitives, and Appendix B varies the utility and necessity margin. These analyses contextualize the outcome-blind proxies; the following section tests what the executable outcome signal itself contributes to supervision selection.

5 OUTCOME-GUIDED SUPERVISION

5.1 EXECUTABLE PREFIXES IDENTIFY HIGHER-YIELD GUIDANCE

The audit measures a directional benefit that can also inform where to apply teacher targets. We examine *Horizon-Split Necessity Gating* (HSNG): estimate necessity from a short executable prefix, then evaluate the selected subset using the terminal outcome. Let $\widehat{U}_{16}(s,a)$ use the same success/progress weights as the terminal utility, but evaluated at horizon 16. The prefix score and selection rule are

$$\widehat{N}_{16}(s) = \sum_{a \in \mathcal{A}(s)} \big(q_T(a \mid s) - p_S(a \mid s)\big)\widehat{U}_{16}(s,a), \qquad g_{16}(s) = \mathbf{1}[\widehat{N}_{16}(s) > 0.01]. \tag{7}$$

The experiment tests whether this score identifies states with positive terminal guidance benefit, and whether the selected subset has positive mean terminal gain. It reuses the 120 ALFWorld and 100 WebShop states; weights and terminal labels follow the audit protocol.

Table 2: **Prefix outcomes enrich beneficial guidance.** State-pair-weighted results on the frozen populations; brackets denote 95% task-cluster bootstrap intervals. Accepted gain is the terminal Teacher-minus-Student utility difference among states with $\widehat{N}_{16} > 0.01$.

Metric	ALFWorld	WebShop
States / task clusters	120 / 71	100 / 73
H16 selected states / selected clusters	19 / 17	36 / 32
H4 AUROC	0.5643	0.5622
H8 AUROC	0.6768	0.6352
H16 AUROC	0.8595 [0.7286, 0.9579]	0.8693 [0.7796, 0.9419]
H16 accepted terminal gain	0.2082 [0.0631, 0.3556]	0.1017 [0.0570, 0.1545]

Table 3: **Frozen H16 on new tasks with independent rollout evaluation.** The new states and task clusters are disjoint from the exploratory populations. Prefix and terminal rollouts are independent. Brackets denote 95% task-cluster bootstrap intervals. Accepted gain is the mean terminal Teacher-minus-Student conditional utility difference on states accepted by the fixed H16 rule. It is a conditional utility quantity.

Metric	ALFWorld	WebShop
New states / task clusters	80 / 40	80 / 40
Candidate-support coverage	100%	100%
Technical failure records	0	0
H16 selected states / task clusters	16 / 13	24 / 18
H16 AUROC	0.80	0.79
AUROC 95% task-cluster CI	[0.68, 0.89]	[0.67, 0.88]
Accepted terminal gain	$+0.10$	$+0.07$
Gain 95% task-cluster CI	[0.04, 0.16]	[0.02, 0.12]

Table 2 shows that H16 discriminates beneficial guidance more clearly than the shorter prefixes. Its accepted-gain intervals lie above zero in both environments. At H8, the ALFWorld accepted-gain point estimate is still -0.0135, showing that the prefix length affects both ranking and the selected subset. Primary prefix and terminal outcomes share trajectories; complementary rollout-slot and top-action analyses are reported in Appendix C.

Independent validation of the frozen H16 rule. We additionally evaluate the fixed rule in Eq. 7 on disjoint new tasks, using independent prefix and terminal rollouts. Each environment contributes 80 new states from 40 new task clusters, with 100% candidate-support coverage and no technical failure records. Table 3 reports AUROCs of 0.80 and 0.79, with positive accepted terminal gains of $+0.10$ and $+0.07$. The favorable estimates remain positive under task-cluster bootstrap intervals. The new-task evaluation remains separate from Table 2 and provides an independent check of the fixed rule.

5.2 Selection information exceeds teacher-target count

Target-level analysis compares selection rules without parameter updates. To isolate the value of selecting useful states from the effect of using fewer teacher targets, we compare H16 with a uniform random gate that selects the same number of states. For a gate g, define the target-level conditional-utility shift

$$G(g) = \frac{\sum_s \omega_s g(s) N(s)}{\sum_s \omega_s}. \tag{8}$$

If k of n states are selected uniformly without replacement, its analytic expectation is $(k/n)\sum_s \omega_s N(s)/\sum_s \omega_s$. We also stratify by initial Teacher–KL quantiles and match the H16 selection count within each bin.

H16 has higher target-level utility than the equal-count random expectation in both frozen populations. The positive point difference remains after matching initial Teacher–KL quantiles, showing that H16

Table 4: **Target-level utility under equal selection counts.** Values use the state-pair weights of the original frozen populations. The uniform-random column is the analytic expectation; the final row gives H16 minus that expectation with a paired task-cluster bootstrap interval.

Metric	ALFWorld	WebShop
Selected states / total states	19 / 120	36 / 100
H16 target-level utility $G(g)$	0.032213	0.036873
Equal-count uniform-random expectation	-0.003713	0.000973
Initial KL-stratified random expectation	-0.003154	0.001458
H16 minus uniform expectation, 95% CI	0.035926 [0.012253, 0.063119]	0.035900 [0.023282, 0.048696]

Table 5: **Accepted learning and off-selection distribution drift after matched micro-training.** Both systems process the same 85 states with six selected states. Improvement is measured against the teacher target; unselected-state drift is measured against the initial student. Larger target improvement and smaller drift are preferred. The comparison reports one training seed; p95 is the nearest-rank quantile over 79 states.

Metric	Exact OPD	HSNG
Selected states with positive target movement	6 / 6	6 / 6
Selected median target-KL improvement ↑	1.8406	1.1723
Unselected mean KL drift ↓	0.35602	0.03345
Unselected p95 KL drift ↓	1.72979	0.17258

captures selection information beyond target count and initial disagreement. Cluster-deletion and complementary rollout-slot analyses are reported in Appendix C.

5.3 Matched updates retain learning with less off-selection drift

We next measure how outcome-guided targets change the student distributions. An ALFWorld training cohort, separate from the audit population, contains 85 states with available H16 records across 45 task clusters, including six selected and 79 unselected states. The six selected states span three task clusters, so this is small-scale mechanism evidence. Exact OPD uses the teacher target at every state. HSNG uses

$$r_s(a) = g_{16}(s)q_T(a \mid s) + \big(1 - g_{16}(s)\big)p_{S,0}(a \mid s), \tag{9}$$

where $p_{S,0}$ is the initial student's frozen distribution on the same candidate set. Unselected states retain their student-anchored targets throughout training. Both runs start from the same Qwen3-1.7B snapshot and use seed 42, the same state and candidate order, three epochs, and 18 optimizer steps. The processed prompt and candidate tokens are identical. Appendix C.3 gives the recipe and accounting.

For selected states, target-KL improvement is the reduction in $D_{\mathrm{KL}}(q_T \| p_S)$ from initialization to the final checkpoint. For unselected states, drift is $D_{\mathrm{KL}}(p_{S,0} \| p_{S,\theta})$. Both quantities are measured on the frozen candidate support with natural logarithms.

HSNG reduces mean unselected-state drift by 90.60%, with a paired task-cluster bootstrap 95% interval of [84.44%, 93.59%]. The p95 point estimate decreases by 90.02%. All six selected states move toward their teacher targets under both recipes. HSNG's median target-KL improvement is 63.7% of Exact's median, showing the change in accepted-target learning magnitude alongside the drift reduction.

Taken together, the two experiments connect an informative outcome signal to an observed training consequence. The prefix study identifies higher-yield teacher targets; the matched-update study quantifies how the target-selection and frozen-student anchoring recipe balances selected-state learning against off-selection distribution change. The two runs process the same states, candidates, tokens, and update steps.

Table 6: **Conditional utility shifts after matched micro-training.** Both checkpoints are evaluated on the same candidates and fixed continuation utilities. Values are arithmetic state means with 95% task-cluster bootstrap intervals.

Evaluation subset or metric	Exact OPD	HSNG
Selected states with positive utility shift	6 / 6	6 / 6
Selected states: mean ΔV_θ	0.144375 [0.003661, 0.213705]	0.025306 [0.001228, 0.037155]
All 85 states: mean ΔV_θ	0.017292 [−0.001481, 0.040588]	0.003049 [−0.000012, 0.006925]

5.4 Target movement and conditional utility shift

Moving toward the teacher target is evaluated separately from KL. Keeping the original candidate actions and their continuation utilities fixed, we rescore each final checkpoint and define

$$\Delta V_\theta(s) = \sum_{a \in \mathcal{A}(s)} \big(p_{S,\theta}(a \mid s) - p_{S,0}(a \mid s)\big) U(s,a). \tag{10}$$

This measures the conditional utility shift induced by the changed candidate probabilities under the fixed outcome model.

Both recipes produce positive conditional utility shifts on all six selected states, with larger shifts under Exact OPD. Together with Table 5, this shows the trade-off between lower off-selection drift and smaller selected-state learning magnitude; the full-cohort means have intervals that cross zero.

6 Related Work

Selective supervision in agent OPD. DAgger, GKD-style on-policy distillation, and expert corrections place expert feedback on student-visited states or generated sequences, addressing mismatch between training data and student behavior (Ross et al., 2011; Agarwal et al., 2024; Lauffer et al., 2026). Selective methods allocate supervision by action discrepancy or uncertainty (Tan et al., 2026), curricula or phase changes (Li et al., 2026), gold-action feedback (Sun et al., 2026), hindsight error localization (Yeo et al., 2026), or state matching (Liu et al., 2026). These selectors target placement. Whether the teacher has an advantage at a state is a separate question from end-to-end performance because compatible contexts can yield similar outcomes along different paths. We test the correspondence between selector signals and utility under fixed populations, candidate supports, and shared continuation.

Outcome-aware behavioral evaluation. Interactive benchmarks admit multiple executable paths between a state and an outcome (Shridhar et al., 2021; Yao et al., 2022; Wang et al., 2022), so action identity is only one view of behavioral difference. FutureBridge-OPD inserts a teacher bridge at high-disagreement states, uses paired frozen-student continuations, and accepts guidance from changes in teacher-preferred token share (Chen et al., 2026). BNA instead compares full teacher and student distributions on a frozen executable candidate support, defines utility by success and progress, and separates directional gain from outcome geometry. Prior work uses expert cost-to-go (Ross & Bagnell, 2014), bisimulation metrics (Ferns et al., 2004), and optimal transport (Cuturi, 2013). BNA joins these views by measuring action identity distance, multi-horizon outcome distance, and expected utility difference as separate quantities.

From measurement to updates. TRPO-like methods constrain overall distributional change (Schulman et al., 2015). Our analysis asks how update pressure should be distributed across states. H16 uses a short executable prefix to estimate conditional utility, with its frozen rule checked on disjoint follow-up tasks. HSNG applies teacher targets only to selected states and anchors the others to the initial student. Matched-count and matched-anchoring controls assess placement through selected-state learning, off-selection drift, and conditional utility. BNA, H16, and HSNG connect these audit metrics to supervision allocation while retaining task-level evaluation.

7 Limitations and Scope

The primary audit uses the Qwen3-1.7B/Qwen3-32B pair, frozen executable candidate supports, the specified stateful continuation policy, and the utility in Eq. 5. Paired rollouts keep the continuation policy and randomization aligned across candidates. The formal populations contain 120 ALFWorld and 100 WebShop states, so the reported estimates refer to these populations and this protocol.

ScienceWorld is not included in the population audit: two restoration checks each recovered 98/100 states, so the preregistered gate stopped before candidate-support construction and outcome-geometry analysis. The gate record is given in Appendix G. The concurrent-method study evaluates reconstructed selection primitives under the same frozen protocol. The exploratory H16 analysis uses the audited populations and shares trajectories between prefix and terminal measurements; complementary slot and cluster analyses appear in Appendix C. A separate frozen-rule follow-up applies H16 to disjoint new tasks with independent rollouts and evaluates selector transfer. The parameter-update evidence comes from a single-seed, 85-state mechanism study, with six selected states spanning three task clusters, and measures candidate-distribution change under the specified Exact and HSNG recipes.

8 Conclusion

On-policy distillation has long treated teacher–student disagreement as the primary signal for where supervision is needed. This paper shows that disagreement and supervision value are different things. We introduce BNA to make the distinction precise: by measuring action-distribution distance, executable outcome distance, and directional utility gain on a shared frozen candidate set, BNA reveals that large surface disagreement coexists with much smaller outcome separation, and that beneficial guidance is concentrated at a minority of states. H16 shows this minority can be identified cheaply from a short executable prefix, and HSNG shows that restricting teacher targets to those states keeps supervision where it helps while protecting the rest of the policy from unnecessary drift.

The practical implication is that supervision placement matters as much as supervision quantity. Methods that allocate teacher pressure based on disagreement alone are spending most of that pressure where it does not help. BNA provides a protocol for measuring this directly, and HSNG offers a tractable mechanism for acting on it.

Ethics statement

This work studies supervision allocation in language-model agents operating on public interactive benchmarks (ALFWorld, WebShop). No human subjects are involved and no new datasets of personal data are collected or released. The student and teacher models are general-purpose language models evaluated on task-completion benchmarks.

AI use statement

Generative AI tools were used for coding assistance, experiment monitoring, and light prose editing. All research questions, experimental designs, and scientific claims were developed by the authors. Artifact hashes and numerical results were verified by the authors against original outputs.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation, 2026. URL https://arxiv.org/abs/2608.01953.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300. Curran Associates Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pp. 950–951. AAAI Press, 2004.

Niklas Lauffer, Xiang Deng, Srivatsa R Kundurthy, Brad Kenstler, and Jeff Da. Imitation learning for multi-turn LM agents via on-policy expert corrections. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=l0LA4SUA50.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents, 2026. URL https://arxiv.org/abs/2606.15912.

Junzhuo Liu, Weiwei Li, Jun Ling, and Peng Wang. When privileged guidance misaligns: State-matched routing and contextualized self-distillation for multi-turn agents, 2026. URL https://arxiv.org/abs/2608.05219.

Stephane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning, 2014. URL https://arxiv.org/abs/1406.5979.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning, 2021. URL https://arxiv.org/abs/2010.03768.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Bowen Song, Weiqiang Wang, Zilei Wang, and Liang Wang. GAPD: Gold-action policy distillation for agentic reinforcement learning in knowledge base question answering, 2026. URL https://arxiv.org/abs/2605.29584.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. ATOD: Annealed turn-aware on-policy distillation for multi-turn agentic tasks, 2026. URL https://arxiv.org/abs/2606.27814.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, 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. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.775/.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

Woongyeng Yeo, Yumin Choi, Taekyung Ki, and Sung Ju Hwang. HINT-SD: Targeted hindsight self-distillation for long-horizon agents, 2026. URL https://arxiv.org/abs/2605.17873.

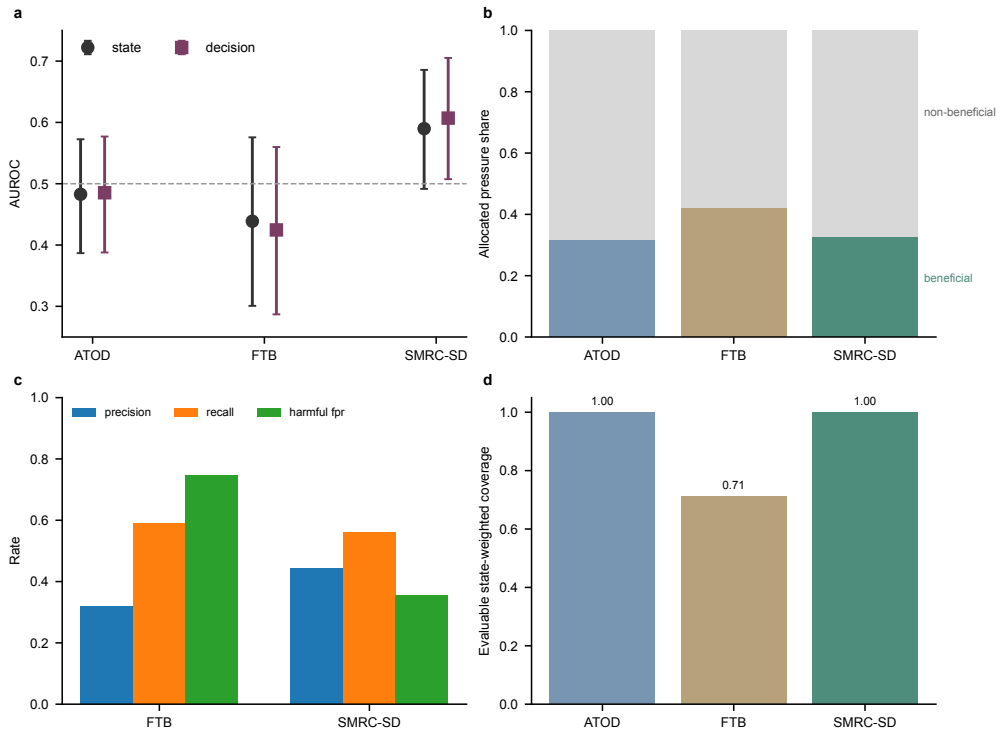


Figure 3: **WebShop selector-signal audit.** (a) Task-cluster AUROC intervals for state and decision estimands. (b) Allocated pressure is split into beneficial and non-beneficial states. (c) Binary-selector precision, recall, and harmful-state false-positive rate. (d) FTB is conditional on 71% state-weighted coverage; technical failures remain in the population with zero allocation.

Table 7: **Bounded WebShop selector reconstructions.** AUROC is state weighted unless marked decision; gain is the state-weighted mean $N(s)$ among accepted states.

Signal	Coverage	AUROC	Decision AUROC	Accepted gain
ATOD re-tokenized	100%	.483 [.387,.573]	.485 [.388,.577]	continuous
FTB H=3	71%	.439 [.301,.576]	.425 [.287,.560]	$-.0029$ [$-.046$, .044]
SMRC-SD match-only	100%	.590 [.492,.686]	.607 [.507,.705]	.0295 [$-.023$, .081]

A Do Concurrent Selector Primitives Help?

We next ask whether three concurrent OPD selector primitives better identify WebShop's beneficial subset, supplementing the likelihood-proxy analysis in Section 4. This is a bounded signal audit, not a reproduction of complete training systems. ATOD uses the released turn-aware difficulty/divergence formula but canonically re-tokenizes historical response text because sampled token IDs were not stored. FTB uses the released future-preference formula under our common H=3 controller; 29 states retain explicit technical failures, leaving a 71-state conditional estimand. SMRC-SD applies the released privileged goal/state matching rule, without reproducing its contextualized self-teacher generation. Workers construct signals without outcome labels; labels enter only at merge.

ATOD and conditional FTB are near chance. SMRC-SD is the strongest signal: its decision-weighted AUROC interval lies above 0.5, while its state-weighted interval touches chance and the accepted-gain interval crosses zero. Its beneficial precision and recall are 0.443 and 0.560, and 67.37% of allocated state pressure remains non-beneficial. Under the common reconstruction protocol, SMRC-SD provides modest discrimination.

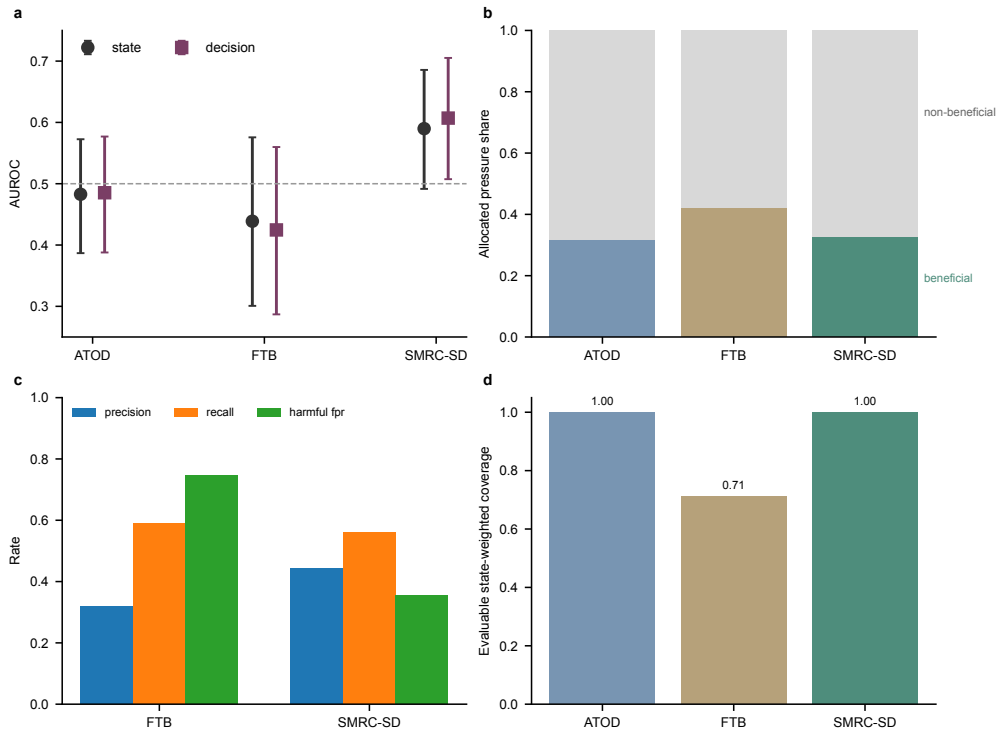


Figure 4: **WebShop selector-signal audit.** (a) Task-cluster AUROC intervals for state and decision estimands. (b) Allocated pressure is split into beneficial and non-beneficial states. (c) Binary-selector precision, recall, and harmful-state false-positive rate. (d) FTB is conditional on 71% state-weighted coverage; technical failures remain in the population with zero allocation.

B Construct Sensitivity

Guidance necessity depends on the utility definition. Before evaluating the selector results, we prespecified a 5×3 grid: success weight in $\{1, .75, .5, .25, 0\}$, progress weight as its complement, and margin in $\{0, .01, .05\}$. Every setting retains all 100 WebShop states and reuses the same candidate outcomes and signals. Equal success/progress weights with $m = 0.01$ define the primary construct; the remaining settings assess sensitivity to these choices.

The label definition affects the results (Figure 5). Minimum agreement with primary labels is 0.370; minimum selector-ranking Kendall τ is 0.333; and the largest change in an applicable pressure metric is 0.249. Under success-only utility, beneficial share falls to about 9–10% and ATOD ranks above SMRC-SD. Under the other 12 settings, SMRC-SD ranks first. Across this grid, no selector wins in every setting. Likelihood disagreement alone does not reliably identify a utility-dependent behavioral-necessity construct. For FTB, stability is estimated on the 71% of states with valid coverage.

C HSNG Measurement and Training Details

C.1 Weighting, horizons and dependence

The prefix analysis reuses the frozen action continuations and adds no model or environment calls. Its primary state-pair weight is the stratum population divided by the sampled stratum count. Decision weights multiply that weight by state multiplicity. Each primary interval uses 10,000 task-cluster bootstrap resamples over 71 ALFWorld or 73 WebShop clusters. The H16 intervals in Table 2 use the stored primary estimator throughout.

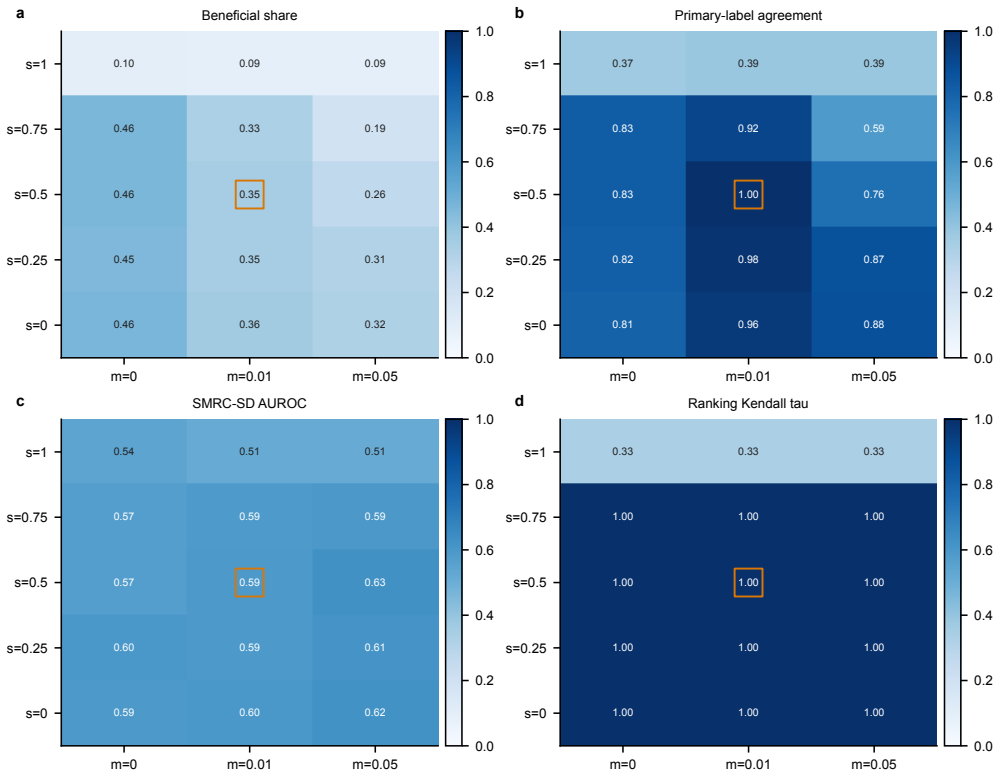


Figure 5: **WebShop construct sensitivity.** Rows vary the success weight s and columns the necessity margin m; the orange box marks the primary construct. (a) Beneficial share, (b) agreement with primary labels, (c) SMRC-SD AUROC, and (d) selector-ranking Kendall τ.

The primary score and terminal label reuse the same four rollout slots. A complementary analysis assigns two slots to the prefix score and the other two to the terminal endpoint, yielding six directed partitions per environment. A sampled-only analysis removes the greedy slot and uses the six nonempty proper subsets of the remaining three slots as prefix sets, with complements used for terminal evaluation. These are correlated partitions, not independent experiments or new-task validation. The four-slot state-pair accepted-gain estimates are all positive; the smallest ALFWorld value is 0.0586. WebShop's worst-partition AUROC is 0.5731, and one ALFWorld decision-weighted accepted-gain estimate is -0.0132. The decision-weighted findings therefore remain distinct from the primary state-pair estimates.

For the four-slot 2/2 partitions, positive point-estimate counts are 6/6 (ALFWorld state-pair), 5/6 (ALFWorld decision), 6/6 (WebShop state-pair), and 6/6 (WebShop decision); the corresponding gain ranges are [0.05856, 0.37593], $[-0.01321, 0.34012]$, [0.02864, 0.08375], and [0.02667, 0.08454]. For sampled-only partitions, the counts are 4/6, 2/6, 6/6, and 6/6 for the same environment/weighting order, with ranges $[-0.10881, 0.31969]$, $[-0.12985, 0.31949]$, [0.01230, 0.09884], and [0.00955, 0.09786].

Using full candidate distributions rather than the H16 top-action contrast increases AUROC by 0.0809 on ALFWorld and 0.0188 on WebShop. We retain distribution-level scoring as the audit's common estimand, while reporting this comparison separately from the value of short-horizon outcome information.

For the target-level comparison in Table 4, uniform random selects exactly the observed number of states without replacement, while the KL-stratified reference matches that count within each initial Teacher–KL quantile bin. Both expectations are analytic; saved mask simulations are checks on the allocation distribution, not training runs or additional observations. Cluster bootstrap recomputes the selected fraction and weighted gain before contrasting H16 with the uniform-random expectation.

C.2 Environment-step accounting

The truncated/full branch counts are 41,022/108,782 environment steps for ALFWorld and 32,986/67,384 for WebShop. These ratios quantify the horizon-truncation cost on recorded branches. They are not wall-clock speedups or all-in training-cost ratios; model scoring, state collection and optimization are separate accounting items.

C.3 Matched micro-training recipe

The source training batch contains 128 states. Schedule-only availability filtering retains 85 states with H16 evidence, across 45 task clusters. The gate selects six states and leaves 79 with frozen-student targets; the six selected states span three task clusters. Both variants start from the same initial checkpoint and use canonical independent-candidate scoring with frozen base calibration. The actual recipe is BF16 full-parameter fine-tuning with AdamW, learning rate 10^{-6}, weight decay 0.01, gradient accumulation over 16 states, gradient clipping at 1.0, and three epochs.

The matched input/update ledger contains 255 state exposures, 1,374 candidate sequences, 211,293 prompt tokens, 4,584 candidate tokens and 18 optimizer steps per variant. There are no extra anchor sequences; nevertheless, Eq. 9 explicitly anchors unselected states to their initial-student distributions through their training targets. The matched design fixes the input and update ledger, not the number of teacher targets, the gradient magnitude, or total upstream environment-evidence cost. The target-level random comparison is not a third trained arm.

C.4 Candidate-distribution measurements

Both final checkpoints are rescored on all 85 states. For selected states, the reported improvement is

$$I(s) = D_{\mathrm{KL}}(q_T \| p_{S,0}) - D_{\mathrm{KL}}(q_T \| p_{S,\theta}) \,. \tag{11}$$

For unselected states, the reported drift is

$$D(s) = D_{\mathrm{KL}}(p_{S,0} \| p_{S,\theta}) \,. \tag{12}$$

Distributions in these equations are normalized on the frozen $\mathcal{A}(s)$. The unselected mean is the arithmetic mean over the 79 states. For ordered drift values $d_{(1)} \le \cdots \le d_{(n)}$, p95 is $d_{(\lceil 0.95n \rceil)}$, which is the 76th value for $n = 79$. It is not an interpolated percentile.

The accepted-learning retention of 63.7% is the ratio of the two accepted medians, $1.1723443/1.8406343$; it is not the median of per-state retention ratios. The recorded paired task-cluster bootstrap uses 10,000 resamples. The Exact-minus-HSNG mean-drift difference is 0.32257 with 95% interval [0.12186, 0.54903]. The HSNG-minus-Exact accepted-median difference is -0.66829, with interval $[-2.40686, -0.66829]$ over 9,550 valid resamples; the remaining resamples contain no accepted cluster. These intervals quantify resampling variation within the observed cohort, while the two training runs each use a single seed.

D Environment-Cost Construction

For each candidate pair (a_i, a_j), the primary environment cost is

$$c_{\mathrm{env}}(a_i, a_j) = \tfrac{1}{4} c_{\mathrm{valid}} + \tfrac{1}{4} c_{\mathrm{success}} + \tfrac{1}{4} c_{\mathrm{progress}} + \tfrac{1}{4} c_{\mathrm{time/failure}}. \tag{13}$$

Each family is normalized to $[0, 1]$ within a task family before averaging across observed horizons. Right-censoring and observed terminal failure are distinct records. Candidate traces retain action validity, observation and state hashes, horizon outcomes, continuation identity, and cost ledgers. The transport solver uses $\epsilon = 0.05$ and explicit marginal checks. Identity, normalized edit, frozen teacher hidden-state cosine, immediate progress, and terminal marginal-gain costs are retained as diagnostics.

E Additional Population Details

ALFWorld's 120 states cover all frozen task strata and have 100% candidate support coverage. Student captured mass has minimum/median/mean 0.9937/1.0000/0.9999; teacher captured mass

is 0.8416/0.9999/0.9956. The Oracle stage uses 129,530 environment steps and 126,243 policy generations including audits. WebShop uses 81,616 environment steps, 80,970 generations, 147.6M prompt tokens, and 1.0M completion tokens.

For WebShop, decision weighting closely matches the state estimand: beneficial/unnecessary/ harmful shares are 35.01/23.29/41.70%; mean teacher gain is 0.00238 with interval $[-0.03448, 0.03880]$. For ALFWorld they are 19.64/68.09/12.27%; mean gain is -0.00435 with interval $[-0.10418, 0.07719]$.

F SELECTOR FIDELITY AND FAILURE TAXONOMY

ATOD retains all turns in the 73 task clusters represented by the selected WebShop states, then applies the released trajectory-level normalization and continuous SOFT-OR formula. Historical generation token IDs were unavailable, so decoded text was canonically re-tokenized; this boundary is included in the method name and artifact manifest.

FTB has 29 technical failures: 25 `invalid_continuation_action`, two `full_continuation_unavailable`, and two records containing two invalid continuation failures. No state is removed. Technical records receive zero allocation and remain in the coverage denominator. The evaluable state-weighted and decision-weighted coverage is 0.7124 and 0.7142. SMRC-SD and ATOD have complete coverage.

G SCIENCEWORLD RESTORE GATE

ScienceWorld (Wang et al., 2022) was preregistered as a third environment, with 100 states across 30 task families. Collection and teacher scoring completed, but strict snapshot restoration did not. The formal primary audit restored 98/100 states after bounded retry; a second independent census also restored 98/100, with different failed state IDs. Because the gate required exact 100/100 restoration, we did not run candidate support, outcome geometry, necessity analysis, or selector evaluation. The four-state engineering pilot is not included in any population claim.