HIVE: COUNTERFACTUAL HUMAN OVERSIGHT ALLOCATION FOR LONG-HORIZON AGENT WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI agents support a growing share of operational workflows, organizations gain parallel throughput while human reviewers supervise a growing stream of actions. We study how to allocate a limited review budget to workflow nodes where intervention can prevent downstream failures and preserve reviewer capability for future cases. We introduce HIVE, a zero-preserving estimator trained on paired replays. HIVE encodes the factual execution and a label-free hypothetical review, then propagates their semantic difference through a typed dependency graph. We combine it with a practice-dependent competence state and the finite-horizon HIVE-DP planner, which jointly accounts for immediate protection, remaining capacity, and competence transitions. On a reproducible enterprise-workflow benchmark with paired potential outcomes, HIVE improves capacity-constrained allocation over outcome and confidence baselines. Across 16-period episodes, HIVE-DP reduces regret over one-step allocation by 48.4–51.3% at capacity four and 32.0–46.3% at capacity six, with gains that persist under faster decay, weaker retention, and an unmodeled nonlinear retention transition. A held-out transition grid and validation-selected risk controller assess sensitivity to dynamics and estimator disagreement. Together, HIVE's intervention-attributable graph score and HIVE-DP's explicit reviewer state form an auditable estimation–planning interface for scarce, dynamically changing oversight.

1 INTRODUCTION

LLM agents increasingly execute multi-step processes: they reconcile expenses, route service requests, prepare procurement records, and send external messages. Tool-using systems interleave reasoning with environment actions (Yao et al., 2023), while enterprise benchmarks now exercise agents on stateful applications, policy-governed service interactions, and common knowledge-work tasks (Trivedi et al., 2024; Yao et al., 2024; Boisvert et al., 2024; Xu et al., 2025).

This shift changes the control problem. The number of actions produced by the automated layer can grow much faster than the number of people retained to inspect them. A supervisor who once followed one process may now be responsible for dozens of concurrent runs. The relevant question extends beyond eventual task success: a human still has to approve an external communication, resolve ambiguous evidence, or stop an irreversible update, but can do so for only a small fraction of the actions in the queue. The scarce resource is review attention, together with the authority and accountability that accompany it. This allocation perspective complements the established view of human oversight as an active component of an automated system rather than a passive fallback (Bainbridge, 1983).

The allocation of this attention matters. An intervention at an early state-changing node may prevent an error from reaching many descendants; the same effort at a late read-only node may provide little protection. Reviewing every action would also restore much of the latency and labour cost that automated execution is intended to reduce. We consequently study a node-level decision: the expected change in downstream workflow utility when a human (or a human-supervised reviewer agent) is asked to accept, revise, or block a proposed action. The decision must be made from the pre-review trace and under a fixed attention budget.

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.

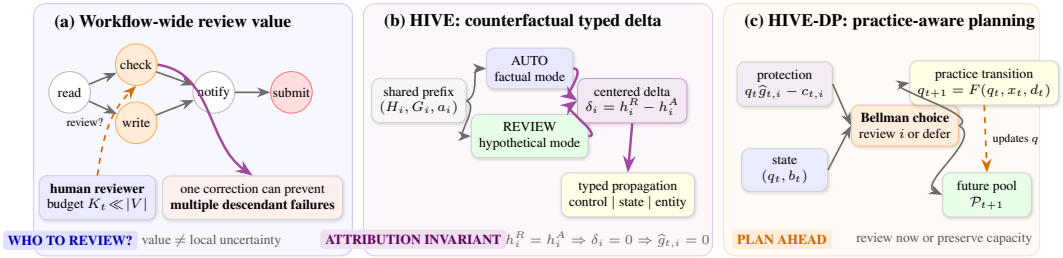


Figure 1: HIVE couples graph-structured counterfactual estimation with practice-aware planning. (a) A scarce reviewer must choose among actions whose errors can propagate through an executable workflow. (b) From the same pre-action state, HIVE contrasts factual AUTO and label-free hypothetical REVIEW semantics, then transports their centered difference over typed descendant relations; the intervention path is exactly zero when the modes coincide. (c) HIVE-DP combines the resulting protection estimate with observed competence and remaining budget, accounting for how review practice changes capability available to future workflow pools.

Review practice creates a second resource effect. Expertise is maintained through repeated exposure; as automated execution handles more routine cases, reviewers encounter fewer opportunities to remain familiar with edge conditions. Prior studies connect reduced active engagement with diminished situation awareness and intervention readiness (Endsley & Kiris, 1995; Bainbridge, 1983). We represent this effect with an observable competence state q_t:

$$q_{t+1} = \min\{1, q_{\min} + (q_t - q_{\min})e^{-\lambda d_t} + \eta x_t(1 - q_t)\}, \tag{1}$$

where d_t is idle exposure and x_t records whether the reviewer practised at time t. The parameters λ and η describe decay and retention. Recent practice initializes q_t, which is observed before scheduling and updated after the allocation. A review can consequently protect the current workflow and preserve part of the capability needed for later workflows.

We present HIVE (Human Intervention Value Estimator), an intervention-attributable estimator and planning interface for this setting. HIVE encodes the factual workflow, constructs a pre-decision hypothetical review representation, and propagates its difference from factual execution through typed descendants of the candidate node. The centered readout is zero when the two representations coincide, isolating the score attributable to the intervention from surrounding workflow difficulty. HIVE is trained with value regression, pairwise ranking, and a differentiable capacity-utility objective. For sequential allocation, its output is separated into a competence-normalized protection gain and an observable review overhead; the resulting planner carries both competence and remaining capacity in a finite-horizon Bellman recursion. We use four names consistently: **HIVE** is the static zero-preserving typed-delta estimator; **HIVE-S** augments it with practice and workflow-local competence features; **HIVE-L** adds forecast-horizon and future-review-pressure features; and **HIVE-DP** schedules HIVE-L protection estimates with an explicit competence state and remaining budget.

Figure 1 makes the interface explicit. The learning problem estimates the change caused by reviewing one candidate from the same pre-action state, distinct from predicting overall workflow difficulty. The planning problem spends those estimated gains while accounting for remaining attention and the effect of review practice on future competence. Under tight capacity, the decisive quantity is the ordering of the few high-value nodes, which need not follow average prediction error.

This interface differs technically from adjacent formulations. Learning-to-defer (Madras et al., 2018; Mozannar & Sontag, 2020) routes examples to experts; HIVE attributes a review's continuation effect to an internal workflow node. Graph treatment-effect estimators (Harada & Kashima, 2022) model structured interference; HIVE imposes zero effect when REVIEW and AUTO semantics coincide and restricts propagation to executable descendants. Workload-aware human–AI allocators (Alves et al., 2024) allocate scarce service; HIVE supplies an intervention score whose reference-state decomposition is compatible with competence-aware planning. These constraints hold for every parameter setting, while unrestricted outcome models must learn them from data.

The paper makes three contributions. (1) We formulate selective human review in agent-operated workflows as a paired potential-outcome problem with a practice-dependent competence transition.

(2) We develop a zero-preserving, typed-delta estimator and a finite-horizon allocator that connect downstream protection with future reviewer capability. (3) We provide an auditable evaluation protocol built from paired replays, typed graphs, workflow-level capacity constraints, held-out structural shifts, and transition stress tests.

Across a two-source enterprise-workflow suite grounded in the operational task requirements of company-work and policy-governed service benchmarks, HIVE improves capacity-constrained allocation over the evaluated baselines, and HIVE-DP consistently improves on one-step scheduling over 16-period episodes. The next section positions the estimator and scheduler relative to selective prediction, graph treatment estimation, and decision-focused learning; Section 3 then gives their common mathematical formulation.

2 Related work

Agent workflows and selective oversight. LLM agents increasingly execute tool-using, state-changing sequences through reasoning–acting loops (Yao et al., 2023), self-reflection (Shinn et al., 2023), and interactive web environments (Zhou et al., 2024). This shift turns human oversight into a scarce control resource and motivates identifying the intermediate intervention that most improves the remaining run. Learning-to-defer methods allocate examples to humans or other experts (Madras et al., 2018; Mozannar & Sontag, 2020), and selective prediction optimizes coverage–risk trade-offs (Geifman & El-Yaniv, 2017). Extensions accommodate multiple experts, workload limits, and human complementarity (Verma et al., 2023; Hemmer et al., 2023; Alves et al., 2024; Wilder et al., 2020; Leitão et al., 2022). These methods supply useful allocation primitives, typically at the level of independent examples. HIVE values a review at an action node by the paired counterfactual change in its executable continuation and allocates reviews under an explicit workflow-level capacity constraint.

Graph structure, heterogeneous effects, and decision-focused learning. Graph representations capture relational dependencies, while dynamic programming supplies the standard decomposition for finite-horizon resource allocation (Sutton & Barto, 2018). Graph-structured treatment estimators such as GraphITE (Harada & Kashima, 2022) and invariant graph-OOD methods such as CIGA and FLOOD (Wu et al., 2022; Liu et al., 2023) demonstrate that structure can support causal or robust prediction. HIVE focuses this structural view on interventions that change a downstream executable continuation. In parallel, meta-learners and quasi-oracle estimators address heterogeneous effects (Künzel et al., 2019; Nie & Wager, 2021), and empirical-welfare and predict-then-optimize methods connect estimates to constrained decisions (Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Wilder et al., 2019; Elmachtoub & Grigas, 2022), including learned decision losses and sequential MDP parameters (Shah et al., 2022; Wang et al., 2021). HIVE combines these concerns through a descendant-restricted semantic delta: the change induced by hypothetical review is transported along typed state, control, and entity relations. HIVE-DP then carries competence and remaining capacity through a Bellman recurrence, leaving the estimator and scheduler independently auditable. When the reviewer state is latent, belief-state learning supplies the standard partially observable extension (Chen et al., 2022); HIVE's scalar recursion is the point-belief case.

Conditional moments, causal paths, and positioning. Conditional moment restrictions (Chamberlain, 1987; Dikkala et al., 2020), varying-coefficient models (Hastie & Tibshirani, 1993), and causal path analysis (Pearl, 2009; Imai et al., 2010) provide established statistical language for conditional and mediated effects. Runtime systems such as WaitGPT (Xie et al., 2024) motivate the intervention interface; HIVE estimates continuation value before review and schedules it under capacity. Together, these perspectives motivate a label-free hypothetical-review representation, a changing reviewer state, and separate evaluations of the graph estimator and stateful allocator developed below.

3 Practice-aware human intervention allocation

We now formalize the two components separated in Figure 1. At each decision time $t = 1, \ldots, T$, the pool $\mathcal{P}_t$ contains candidates indexed by i. Candidate i has a factual history H_i, typed dependency graph G_i, proposed action a_i, and a hypothetical review-mode descriptor m_i. Reviewer competence $q_t \in [0, 1]$ is an observable operational-performance state estimated from recent task-relevant practice. The scheduler selects $x_{t,i} \in \{0, 1\}$ subject to the review capacity $\sum_{i \in \mathcal{P}_t} x_{t,i} \leq K_t$. We write

$x_t = \mathbf{1}\{\sum_{i\in\mathcal{P}_t} x_{t,i} > 0\}$ for period-level review practice; the dynamic allocation studied below admits at most one review per period.

The immediate causal value of review is

$$\Delta_i(q_t) = \mathbb{E}[J_i^A - J_i^{AH} \mid H_i, G_i, a_i, m_i, q_t], \tag{2}$$

where both potential outcomes start from the same pre-action state. The AH branch invokes a fixed review policy that accepts, revises, or blocks the proposal; when reviewers differ, its pre-decision descriptor r_i is part of H_i and the intervention is the policy conditional on r_i. J_i^{AH} includes the resulting action and every subsequent response of the frozen downstream agent. Competence also changes with use. Given elapsed idle exposure d_t and whether a review is performed, we use

$$q_{t+1} = F(q_t, x_t, d_t) = \min\{1, q_{\min} + (q_t - q_{\min})e^{-\lambda d_t} + \eta x_t(1 - q_t)\}. \tag{3}$$

The resulting finite-horizon objective is

$$\max_{\pi}\ \mathbb{E}_\pi\left[\sum_{t=1}^{T}\gamma^{t-1}\left(\sum_{i\in\mathcal{P}_t} x_{t,i}\Delta_i(q_t) - c_i x_{t,i}\right)\right], \tag{4}$$

with state $(q_t, \mathcal{P}_t)$ and policy π. This formulation captures the possibility that a marginal case merits review because its practice preserves competence for a later, high-impact case, an effect absent from static escalation. Mandatory safety reviews are reserved before this optimization: if $\mathcal{M}_t$ is the policy-defined mandatory set, HIVE ranks the remaining pool with residual capacity $\max\{0, K_t - |\mathcal{M}_t|\}$; excess mandatory demand activates the organization's safety escalation protocol, with model ranking applied to the residual discretionary capacity.

Identification in the replay protocol. The causal interpretation follows from state-matched paired replays. It requires consistent interventions, the same restored state and exogenous settings in both branches, no cross-workflow interference, and finite-mean evaluator noise. The paired difference is then unbiased for $\Delta_i(q_t)$, while the REVIEW representation serves as its pre-intervention predictor. Reviewer identity is part of the intervention, so pooled analysis conditions on the pre-decision reviewer descriptor. For settings with logged review and overlap, a cross-fitted doubly robust alternative applies under conditional exchangeability. The experiments use the paired-replay regime.

HIVE estimator. Figure 1(b) corresponds to two coupled encoders. The factual encoder assigns each node v an initial state from its action text, node type, and pre-intervention numeric features, then applies relation-specific messages over control, state, and entity edges. This produces a context state z_v that records what has happened and which later actions depend on it. The intervention encoder is narrower: at candidate i it is seeded only by the difference between frozen REVIEW and AUTO representations,

$$d_i^{(0)} = h_i^R - h_i^A, \qquad d_v^{(0)} = 0 \quad (v \neq i). \tag{5}$$

For layer ℓ, a relation-specific, context-gated message is averaged over incoming edges and then masked to descendants of i:

$$m_v^{(\ell)} = \frac{1}{\max(1, \deg^-(v))}\sum_{r\in\mathcal{R}}\sum_{u:(u,v)\in E_r} W_{\ell,r} d_u^{(\ell)} \odot g_\ell(z_u, z_v), \tag{6}$$

$$d_v^{(\ell+1)} = \mathbf{1}[v \succeq i]\,\mathrm{LN}_0\Big(d_v^{(\ell)} + \mathrm{GELU}(D_\ell d_v^{(\ell)} + m_v^{(\ell)})\Big). \tag{7}$$

where $\mathcal{R}$ is the three-relation vocabulary. Relation-specific weights distinguish, for example, a corrected identifier carried by a state edge from mere control adjacency. The descendant mask restricts intervention effects to the candidate and its downstream executable continuation.

The candidate score sums centered descendant contributions and a separately centered candidate residual,

$$\widehat{B}_i = \sum_{v\succeq i}[\rho(z_v, d_v^{(L)}) - \rho(z_v, 0)] + [\beta(z_i, d_i^{(L)}, n_i) - \beta(z_i, 0, n_i)], \tag{8}$$

where n_i contains pre-review numeric features and LN_0 has no affine offset. The intervention path has no additive bias. Consequently Eqs. 5–8 implement the attribution- invariant strip in Figure 1:

identical AUTO and REVIEW semantics produce exactly zero value, irrespective of workflow length or factual risk. The descendant mask simultaneously guarantees graph locality: $d_v^{(L)} = 0$ for every $v \not\succeq i$. These invariants specify where the model can represent an effect; causal interpretation and graph coverage are supplied by the replay and graph audits. Training combines squared value error, within-pool ranking, and a capacity-aligned relaxed utility with zero-valued dummy actions, allowing the allocator to leave capacity unused when every available review has negative value. A worked procurement case reports the resulting split between local and propagated value.

To obtain an auditable quantity from paired workflow replays, the long-horizon benchmark constructs a reference-state target

$$B_i^+(q_0) = P_i(q_0) - c_i, \tag{9}$$

where $P_i(q_0)$ combines the current candidate's local-plus-downstream protection with its pre-specified horizon interaction. This is one paired label at observed competence q_0, not an additional state transition. HIVE-DP uses the entire gross quantity $P_i(q_0)$ as the reference value for its competence-separable reward and carries the transition F between decision times.

HIVE-S augments HIVE with two pre-intervention candidate features—recent practice frequency and workflow-local competence descriptor s_i—and gates the counterfactual mode delta by s_i. The gate is multiplicative and the readout remains centered, so the original null-intervention invariant is preserved:

$$h_i^R = h_i^A \quad \Longrightarrow \quad \widehat{B}_i^+ = 0. \tag{10}$$

The allocator ranks $\widehat{B}_i^+$ under K_t, as in the static case. A no-skill ablation keeps the six-feature architecture but masks the two competence features, separating information value from parameter count. For the long-horizon variant, two additional pre-review covariates—forecast horizon h_t and future review pressure p_t—are appended. HIVE-L applies the fixed retention interaction $(1+1.5h_tp_t)$ to the delta gate; masking these two columns gives the parameter-matched no-horizon ablation reported with the supplementary one-step results.

Finite-horizon HIVE-DP. HIVE-L's output is converted into an estimated competence-normalized protection gain $\widehat{g}_{t,i}$ and an observable review cost $c_{t,i}$, giving the state-dependent reward

$$\widehat{R}_{t,i}(q) = q\widehat{g}_{t,i} - c_{t,i}. \tag{11}$$

Operationally, HIVE first predicts $\widehat{B}_{t,i}$ from the candidate features, including s_i. The scheduler then anchors that complete score at the observed reference competence q_0 through $\widehat{B}_{t,i} = q_0\widehat{g}_{t,i} - c_{t,i}$. Thus s_i remains part of the candidate value estimate, whereas q is the shared state carried by the controller. In the dynamic experiments, the period-one state satisfies $q_1 = q_0$. This interface allows deployments to use measured latency, staffing, or escalation cost in place of the fixed benchmark overhead while retaining the competence transition. For an observed forecast of arriving workflow pools, HIVE-DP solves with the implementation's admissible review set $\mathcal{P}_t^+(q) = \{i \in \mathcal{P}_t : \widehat{R}_{t,i}(q) > 0\}$,

$$\widehat{V}_t(q,b) = \max\{\gamma\widehat{V}_{t+1}(F(q,0,d_t),b), \tag{12}$$

$$\max_{i\in\mathcal{P}_t^+(q)} \big[\widehat{R}_{t,i}(q) + \gamma\widehat{V}_{t+1}(F(q,1,d_t),b-1)\big]\}, \tag{13}$$

where the inner maximum is omitted when $b = 0$ or $\mathcal{P}_t^+(q) = \varnothing$, b is remaining review capacity, and $\widehat{V}_{T+1} = 0$. In the reported experiment, $\widehat{g}_{t,i}$ is obtained by adding the observable benchmark overhead back to the pre-review HIVE net score and dividing by the episode's observed initial competence; $c_{t,i} = 0.02\,\texttt{review_cost}$. A deployment can replace this constant with measured review costs. We discretize q on an M-point grid, yielding complexity $O(TKM(1+\max_t|\mathcal{P}_t|))$. Thus the paired-replay dynamic experiment makes the following modeling choice explicit:

$$P_i(q) = \frac{q}{q_0}P_i(q_0), \qquad \widehat{R}_{t,i}(q) = \frac{q}{q_0}(\widehat{B}_{t,i} + c_{t,i}) - c_{t,i}. \tag{14}$$

The transition F changes competence for later arriving pools, whereas the learned net score records the current candidate's paired-replay value at the reference state. Reference-state recovery and a finite-horizon error bound hold for the implemented positive-reward action set. The one-step HIVE-L

allocator is recovered by ranking the current B_i^+ values independently at each period; HIVE-DP carries (q, b) across periods. When F is learned, longitudinal logs and cluster-bootstrap parameter intervals define the rectangular transition set used by the robust planner. The coverage gate routes settings outside the calibration envelope to the neutral mandatory-review policy.

Capability option value. Assume $\widehat{g}_{t,i} \geq 0$ and $e^{-\lambda d_t} \geq \eta$. Then $F(q, x, d_t)$ is nondecreasing in q and $F(q, 1, d_t) \geq F(q, 0, d_t)$. Backward induction shows that $\widehat{V}_t(q, b)$ is nondecreasing in competence and remaining capacity. Thus the continuation increment

$$\Omega_t = \gamma[\widehat{V}_{t+1}(F(q, 1, d_t), b-1) - \widehat{V}_{t+1}(F(q, 0, d_t), b-1)] \tag{15}$$

is nonnegative for $b > 0$: HIVE-DP adds a capability option to immediate protection. The shared budget can nevertheless yield different node-wise orderings for dynamic and myopic schedules by reallocating practice across periods. The complete argument also establishes the zero-allocation boundary.

When reviewers specialize by task family, the scalar state generalizes to $\mathbf{q}_t \in [0, 1]^R$ with assignment $a_t \in \{0, 1, \ldots, R\}$. The transition is componentwise, $\mathbf{q}_{t+1} = F(\mathbf{q}_t, a_t, d_t)$, and the reward is $q_{t,a_t}\widehat{g}_{t,i} - c_{t,i}$. A vector-state Bellman recursion searches over feasible reviewer–candidate pairs and has grid complexity $O(TKM^R \max_t |\mathcal{P}_t|R)$; scalar HIVE-DP is the $R = 1$ case used here. A specialization stress test quantifies the cost of collapsing this state to its mean. If competence is partially observed, the state becomes a posterior $\mu_t(q)$ updated from oversight evidence. The corresponding belief-state recursion reduces to Eq. 13 when the posterior is a point mass.

Uncertainty-aware HIVE-DP. With S independently seeded HIVE checkpoints, let $\bar{g}_{t,i}$ and $\sigma_{t,i}$ denote the ensemble mean and standard deviation of the protection estimate. The risk-sensitive planner replaces $\widehat{g}_{t,i}$ by the lower-confidence estimate $\widetilde{g}_{t,i} = \bar{g}_{t,i} - \kappa\sigma_{t,i}$ before solving the same Bellman recursion. Thus model disagreement is converted into a conservative review threshold. We additionally evaluate the recursion under an uncertainty set for competence transitions by using the largest admissible decay and the smallest admissible retention; monotonicity makes this a worst-case transition for future value. Setting $\kappa = 0$ and collapsing the transition set recovers HIVE-DP. The robustification operates at the decision layer while leaving the semantic encoder unchanged. In the calibrated deployment protocol, κ and the transition radius are selected from a small validation grid using a one-standard-error rule: the conservative choice is selected when its validation regret plus one standard error is below the neutral controller. Robustness is thereby an auditable operating-point selection tied to validation evidence.

Intervention-attribution invariants. Because the delta encoder and relation layers are bias-free on the intervention path and both readouts are centered at $\delta = 0$, HIVE satisfies

$$h_i^R = h_i^A \implies \widehat{B}_i = 0 \quad \text{and} \quad v \not\succeq i \implies d_v^{(L)} = 0. \tag{16}$$

These properties are enforced architecturally; review value is therefore attributable to the intervention path, separating it from factual risk and disconnected branches. An unrestricted graph T-learner also represents functions outside these invariants, so, when its unrestricted heads contain the HIVE mapping, its hypothesis class strictly contains HIVE's under identical factual inputs. HIVE's inductive bias is tailored to settings where review effects respect the declared graph. At converging paths, messages aggregate incoming evidence while Eq. 8 counts each descendant once. The corresponding analysis bounds estimation and top-K regret by represented error plus omitted and spurious graph-effect mass. The experiments test the static estimator and its mode/propagation ablations before the competence-aware scheduler and its transition controls.

Appendix map. Appendices A–D give the dynamic, state-error, and identification results; Appendix G contains the complete attribution and allocation proofs, including the graph error bound and belief-state extension. The remaining appendices report the experimental specification, transition learning, specialization and uncertainty analyses, robustness tests, and worked cases.

4 Experiments

Following this decomposition, we first evaluate HIVE's ranking of continuation effects within a fixed pool and then HIVE-DP's scheduling of those estimates across arriving pools. In both cases the

operational question is the one posed in Section 1: when one reviewer can inspect only a fraction of the actions produced by many concurrent workflows, which nodes should receive that attention? The benchmark combines paired factual/counterfactual replays, typed dependency graphs, and workflow-level allocation pools. Its two source contracts are fixed before training from requirements in public company-work and policy-governed service benchmarks, spanning company/service operations, cross-application coordination, policy dialogue, and stateful tools. The constructor varies task identities within these contracts and records every potential outcome.

4.1 Claim, estimand, and information set

Our estimand is the downstream utility change from requesting a review at a candidate node, conditioned on the factual prefix and proposed action. Both branches for candidate i start from the same restored pre-action state and use the same downstream policy. We define

$$B_i = J_i^A - J_i^{AH}, \qquad J = 1 - S + 0.05\frac{N_{\text{step}}}{N_{\max}} + 0.02C_{\text{review}}, \tag{17}$$

where A executes the factual action and AH executes the action produced by the review policy. Both branches then resume the same frozen agent policy from their respective post-action states, so the contrast includes induced changes to the entire continuation. The coefficients are fixed before test evaluation and evaluated in a cost-weight sensitivity analysis. The static estimand contains no reviewer-learning term; that term is introduced explicitly below.

Before requesting review, the allocator observes the task instruction, factual prefix, current observation, proposed action, typed factual graph, numeric node features, frozen AUTO and hypothetical-REVIEW representations, and—in the skill-aware split—recent practice frequency and current competence. The allocation-time information boundary excludes the reviewer's eventual accept/revise/block decision, any revised action, the evaluator score, and the future trace. The paired continuation outcome B_i serves exclusively as a training or validation label.

The `[REVIEW]` pooling prompt describes a label-free policy that may accept, revise, or block the proposed action. Its allocation-time embedding is formed before the realized decision and replay score become available. The review output enters the paired label and replay job.

4.2 Practice-dependent oversight setting

Each workflow instance receives pre-review practice f_i and a workflow-local competence descriptor s_i before label construction. The pre-specified paired target is

$$B_i^{\text{ref}} = s_i\Delta_i + h_i(s_i) - c_i, \tag{18}$$

where Δ_i is local-plus-downstream protection and $h_i(s_i)$ is the pre-specified candidate-local horizon interaction used in benchmark construction; both are fixed before test generation. This equation defines one candidate label, not a second transition law. HIVE observes (f_i, s_i) and gates the mode delta by s_i. In the multi-period experiment, s_i remains a candidate covariate, while the scheduler has a separate episode-level competence state initialized at q_0. The predicted or realized gross reference protection $B_i^{\text{ref}} + c_i$ is anchored at that q_0 and extended by $P_i(q) = (q/q_0)(B_i^{\text{ref}} + c_i)$, exactly as implemented by HIVE-DP. The parameter-matched 'no-skill' ablation masks both state features, isolating this information from model capacity.

4.3 Data and paired replay protocol

The reproducible benchmark contains 240 workflows in 12 families balanced across two benchmark-grounded sources, split by task identity into 144/48/48 train/validation/test workflows and 720 candidate interventions. Source A retains the shared business entities, cross-step state reuse, permission checks, and externally visible actions specified by TheAgentCompany and WorkArena++; Source B retains the policy lookup, approval-gated mutation, missing-information, and stateful tool requirements in τ-bench and ToolSandbox (Xu et al., 2025; Boisvert et al., 2024; Yao et al., 2024; Lu et al., 2025). Within each contract, a deterministic constructor varies task identity, typed resources, evidence ambiguity, permission conditions, and external visibility while instantiating the

required control/state/entity dependencies; every method shares these task-derived graphs. Review value combines local risk, descendant exposure, detectability and review overhead; its equation and parameters are hidden from all models. Validation/test targets average three outcome draws; training uses one.

Each paired replay restores the same pre-action state, executes AUTO or REVIEW, resumes the frozen downstream policy, and records score, steps, review cost, snapshot, and seed. Pairing fixes task, prefix, seed, model, decoding, and step limit; failed replays remain in the audit log under pre-specified exclusions.

A held-out 160-workflow structural-shift set changes depth, detectability, and utility (sublinear risk and delayed exposure saturation).

An additional structural study scales the static evaluation to 1,500 workflows and five source-specific interaction contracts. It preserves the original matched comparison and the dynamic experiments below.

4.4 Models and training

HIVE maps the factual AUTO representation and four numeric features with a relation-aware graph encoder. A separate delta encoder starts at the candidate with the difference between the hypothetical REVIEW and AUTO representations, propagates over graph descendants, and combines those contributions with a candidate-local residual. The construction reflects the hypothesis that a review changes a node's semantics and that the change can travel through downstream state dependencies.

Comparators share candidates and frozen representations. L2D adapts learning to defer (Madras et al., 2018; Mozannar & Sontag, 2020). DeCCaF, FLORA, CAIR, and Twin-Outcome instantiate independent outcome heads, graph treatment, candidate masking, and graph T-learning patterns (Künzel et al., 2019; Harada & Kashima, 2022; Wu et al., 2022). Direct-Policy is a candidate-local plug-in policy learner (Kitagawa & Tetenov, 2018; Athey & Wager, 2021), and Confidence is a fixed heuristic. Free-Delta is a candidate-local MLP over the REVIEW–AUTO difference and numeric features in a candidate-local architecture; its capacity-matched GNN version is retained as an architectural ablation.

HIVE minimises value MSE, a pairwise ranking loss, and a differentiable capacity utility loss. Each batch contains at most one candidate per workflow. Twin-Outcome and FLORA-Marginal receive the identical ranking and capacity losses on their outcome-difference scores; this holds decision-focused supervision fixed while changing the intervention architecture. L2D, DeCCaF, and CAIR retain their native classification or outcome objectives. AdamW, gradient clipping, hidden size 64, three graph layers, and dropout 0.1 are fixed in the paired-replay study. HIVE and every learned baseline in the headline comparison are fitted with seeds 11, 17, 23, 29, 31, 37, and 41; architecture ablations use seed 11 for a focused architecture evaluation. Checkpoints are selected on validation regret, and the same exact allocator is used for validation and test scoring. All learned methods share candidates, frozen representations, optimizer budget, patience, and seeds; graph models also share width and depth. Native objectives and model selection use validation information exclusively; the Free-Delta variants fix their validation pool across epochs, favoring them by removing resampling noise.

Risk-sensitive extension. HIVE-DP uses Section 3's protection/overhead decomposition. A three-checkpoint robust variant substitutes $\widetilde{g} = \bar{g} - \kappa\sigma$. Validation applies the one-standard-error rule to neutral, lower-confidence nominal, and worst-transition candidates; the selected controller is frozen before the held-out transition audit.

4.5 Evaluation and statistical analysis

For a pool of distinct workflows and capacity K, the exact allocator solves

$$\max_{x} \sum_{i} \widehat{B}_i x_i, \quad \sum_{i} x_i \le K, \quad \sum_{i \in V_e} x_i \le 1, \quad x_i \in \{0,1\}, \tag{19}$$

where V_e is the set of candidates belonging to workflow e. The one-candidate-per-workflow constraint defines the primary estimand and isolates each workflow's allocation contribution; a separate stress test quantifies repeated-review behavior. We report MAE, Spearman correlation, calibration/error

Table 1: Seven-seed paired-replay results. Brackets count HIVE wins; Confidence is deterministic. Regret@8 is lower-is-better.

Method	ρ	Source A Regret@8	Source B Regret@8
Confidence	.7698	.1921	.1149
Direct-Policy	.8893	.2093 [7/7]	.1574 [6/7]
L2D	.8418	.2772 [7/7]	.6492 [7/7]
DeCCaF	.3543	.5316 [7/7]	.7003 [7/7]
FLORA-Marginal	.7555	.1739 [6/7]	.2034 [7/7]
CAIR-Mask	-.0666	.7030 [7/7]	1.0010 [7/7]
Twin-Outcome	.8894	.1712 [6/7]	.0715 [6/7]
Free-Delta	.8381	.1899 [6/7]	.1342 [7/7]
HIVE	**.8944**	**.0844**	**.0360**

Table 2: Architecture ablations on the held-out test set, seed 11.

Configuration	MAE	ρ	Source A Regret@8	Source B Regret@8
Full HIVE	.0674	.9268	.0512	.0100
mode semantics removed	.0434	–	.8715	1.1070
candidate-local delta	.0568	.9244	.0565	.0100

Table 3: Dynamic regret under nominal dynamics: mean (SD) over 700 paired episodes per source.

Capacity	Policy	Source A	Source B
$K = 4$	no-horizon static	.0512 (.0399)	.0399 (.0266)
	HIVE-L one-step	.0511 (.0398)	.0403 (.0272)
	HIVE-DP	**.0249** (.0262)	**.0208** (.0152)
$K = 6$	no-horizon static	.0487 (.0339)	.0467 (.0284)
	HIVE-L one-step	.0529 (.0373)	.0403 (.0254)
	HIVE-DP	**.0284** (.0280)	**.0274** (.0183)

by value decile, utility, oracle utility, and regret at $K \in \{4, 8, 12\}$. Each source has 24 held-out workflows, so its headline statistic uses the complete source pool at each seed. Confidence intervals resample task clusters; paired seed differences are reported separately. Cost-weight and capacity sensitivity are mandatory because a single operating point can reverse rankings.

Across seven seeds HIVE has the best mean on all three criteria. It reduces Free-Delta regret by .1055/.0982 on A/B, winning 6/7 and 7/7 seeds ($p = .0088/.0083$), and wins 6/7 against Twin-Outcome on both sources ($p = .0055/.0568$). The paired analysis also reports dispersion and the matched-GNN ablation.

Mode removal produces near-constant scores with weak allocation despite low pointwise error. The candidate-local delta shows that propagation helps Source A while local evidence serves Source B; the full model retains both signals.

On the unseen response/depth/utility shift, seed-11 HIVE retains Regret@8 of $.0244/.0291$ on A/B. Competence reduces regret by 16.2/43.4% against its matched ablation; supplementary results give the full one-step comparison.

4.6 Finite-horizon scheduling

We evaluate 16-period held-out episodes with one review per period at most, $K \in \{4, 6\}$, $q_1 = q_0 \in \{.45, .60, .75\}$, and nominal dynamics $.05/.08/.97$. HIVE-DP uses a 151-point grid; all policies share realized gains, transitions, and rewards across 100 orderings, two sources, and seven seeds.

HIVE-DP reduces regret over one-step HIVE-L by 51.3/48.4% on Sources A/B at $K = 4$ and 46.3/32.0% at $K = 6$; every paired episode difference is positive. The gain persists under faster decay and weaker or nonlinear retention, linking it to timing and competence propagation.

Appendices E–R provide the full protocol, additional results, robustness analyses, and cases.

5 Conclusion

We introduced HIVE, which combines a zero-preserving typed-delta estimator with competence-aware finite-horizon planning for scarce review. The estimator confines AUTO–REVIEW differences to executable descendants, while HIVE-DP carries protection, overhead, competence, and budget. Across paired replays and structural and transition shifts, the interface improves constrained allocation under auditable assumptions. It supports resource scheduling as attention and capability evolve; future work will address heterogeneous teams and temporally delayed workflow benefits.

Ethics Statement

HIVE allocates discretionary review after reserving capacity for legal, safety-critical, or irreversible actions designated as mandatory by policy. In deployment, competence measures should support system-level audits of measurement quality and subgroup effects. Organizations should also monitor workload concentration, reviewer autonomy, and appeal paths.

Reproducibility Statement

The main paper specifies the problem formulation, estimator, planner, evaluation protocol, and principal results. Appendix F records the benchmark construction, data splits, hyperparameters, model selection, allocation, and statistical procedures; Appendix G provides complete proofs, and the remaining appendices report robustness analyses and workflow cases.

AI Use Statement

Generative AI tools were used as assistive tools during manuscript preparation and revision. The authors reviewed and verified the final manuscript and take responsibility for its contents. Generative AI systems are not listed as authors.

References

Jean V. Alves, Diogo Leitão, Sérgio Jesus, et al. Cost-sensitive learning to defer to multiple experts with workload constraints. *arXiv preprint arXiv:2403.06906*, 2024.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 2021.

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. doi: 10.1016/0005-1098(83)90046-8.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. WorkArena++: Towards compositional planning and reasoning-based common knowledge work tasks. *arXiv preprint arXiv:2407.05291*, 2024.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 1987.

Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, and Jianyu Chen. Flow-based recurrent belief state learning for pomdps. In *Proceedings of ICML*, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, 2020.

Adam N. Elmachtoub and Paul Grigas. Smart predict, then optimize. *Management Science*, 2022.

Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, 1995. doi: 10.1518/001872095779064555.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Shonosuke Harada and Hisashi Kashima. Graphite: Estimating individual effects of graph-structured treatments. *Transactions of the Japanese Society for Artificial Intelligence*, 2022.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society B*, 1993.

Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *Proceedings of AAAI*, 2023.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 2010.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 2018.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *PNAS*, 2019.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2025.

Diogo Leitão, Pedro Saleiro, Mário Figueiredo, and Pedro Bizarro. Human-ai collaboration in decision-making: Beyond learning to defer. *arXiv preprint arXiv:2206.13202*, 2022.

Yang Liu, Xiang Ao, Fuli Feng, et al. Flood: A flexible invariant learning framework for out-of-distribution generalization on graphs. In *Proceedings of KDD*, 2023.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL*, 2025.

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

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of ICML*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems*, 2022.

Noah Shinn, Federico Cassano, Ashish Berman, Arnav Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts. In *Proceedings of AISTATS*, 2023.

Kai Wang, Sanket Shah, Haipeng Chen, Andrew Perrault, Finale Doshi-Velez, and Milind Tambe. Learning mdps from features: Predict-then-optimize for sequential decision making by reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of AAAI*, 2019.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of IJCAI*, 2020.

Ying-Xin Wu, Xiang Wang, An Zhang, et al. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022.

Liwenhan Xie, Chengbo Zheng, Haijun Xia, et al. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. In *Proceedings of UIST*, 2024.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2025.

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.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Srivatsan, Zhe Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

PART I: THEORY, IDENTIFICATION, AND GUARANTEES

A MONOTONICITY OF THE DYNAMIC VALUE

We prove the structural claim used by HIVE-DP. Write $P(u) = \min\{1, \max\{q_{\min}, u\}\}$ and let $F_x = P \circ \bar{F}_x$ denote the clipped transition after decision $x \in \{0, 1\}$.

A.1 ASSUMPTIONS AND OPERATIONAL AUDIT

Scope and empirical meaning of the assumptions. The condition $a = e^{-\lambda d} \geq \eta$ says that the competence retained from a marginal increase in current competence is at least the reduction of the practice increment as competence approaches its ceiling. It is directly checkable from transition estimates; the nominal experiment has $a = e^{-1.8(0.05)} = .914$ and $\eta = .08$. Clipping preserves the resulting ordering. The assumption $\widehat{g}_{t,i} \geq 0$ concerns gross protection, before the observable review cost is subtracted. It matches the candidate construction in which review can avert loss; the implemented positive-reward gate still skips candidates whose net value $q\widehat{g}_{t,i} - c_{t,i}$ is not positive. These are therefore properties of the declared operating envelope, not assumptions that every unrestricted predictor output is beneficial.

Table 4: Assumption-to-measure audit. Empirical quantities use seven checkpoints; the residual radius is a finite-sample diagnostic paired with an explicit population-coverage gate.

Condition	Operational quantity	Audit value	Operational extension
$e^{-\lambda d} \geq \eta$	nominal transition-order margin	$.9139 - .08 = .8339$	validation-selected robust transition
$\widehat{g} \geq 0$	nonnegative predicted gross gain	51.8% of candidate/reference rows	net-positive action gate
$\lvert g - \widehat{g}\rvert \leq e^g$	95th-percentile held-out normalized residual	$e^g = .3446$	widen interval or collect replays
gate stability	$\lvert\widehat{R}\rvert > qe^g$ on test candidate–grid states	13.1% satisfy this sufficient certificate	skip/mandatory routing near zero
Lipschitz envelope	$q \in [.25, 1]$, clipped affine transition	slope at most .9139; finite gain range	coverage gate outside range
cluster independence	resampling unit	task-identity workflow cluster	keep all within-workflow nodes together

The gate-margin row deliberately audits the sufficient uniform-sign condition used by Proposition 3. The scheduler operates across the full state space, and the bound applies to states carrying this certificate; the stated fallback covers the complementary region. This reports theorem coverage and empirical performance as distinct, auditable quantities.

Lemma 1 (Ordering of competence transitions)**.** *Fix $d \geq 0$, $\lambda \geq 0$, $\eta \geq 0$, and assume $a := e^{-\lambda d} \geq \eta$. Then (i) $F_x(q)$ is nondecreasing in q for both actions, and (ii) $F_1(q) \geq F_0(q)$ for every $q \in [q_{\min}, 1]$.*

Proof. Before clipping,

$$\bar{F}_0(q) = q_{\min} + (q - q_{\min})a \tag{A1}$$
$$= q_{\min} + aq - aq_{\min} \tag{A2}$$
$$= (1 - a)q_{\min} + aq, \tag{A3}$$
$$\bar{F}_1(q) = q_{\min} + (q - q_{\min})a + \eta(1 - q) \tag{A4}$$
$$= q_{\min} + aq - aq_{\min} + \eta - \eta q \tag{A5}$$
$$= (1 - a)q_{\min} + \eta + (a - \eta)q. \tag{A6}$$

Equations (A2) and (A5) distribute a and η; (A3) and (A6) collect constant and q terms. For $q' \geq q$, subtracting the corresponding affine forms gives

$$\bar{F}_0(q') - \bar{F}_0(q) = [(1 - a)q_{\min} + aq'] - [(1 - a)q_{\min} + aq] \tag{A7}$$
$$= a(q' - q) \geq 0, \tag{A8}$$
$$\bar{F}_1(q') - \bar{F}_1(q) = [(1 - a)q_{\min} + \eta + (a - \eta)q'] - [(1 - a)q_{\min} + \eta + (a - \eta)q] \tag{A9}$$
$$= (a - \eta)(q' - q) \geq 0. \tag{A10}$$

The last inequalities use $a > 0$, $a - \eta \geq 0$, and $q' - q \geq 0$. Next, subtracting (A1) from (A4) cancels the shared decay term:

$$\bar{F}_1(q) - \bar{F}_0(q) = [q_{\min} + (q - q_{\min})a + \eta(1-q)] - [q_{\min} + (q - q_{\min})a] \tag{A11}$$
$$= \eta(1-q) \geq 0, \tag{A12}$$

because $q \leq 1$. Finally, P is nondecreasing: if $u \leq v$, then clipping both numbers to the same interval cannot reverse their order. Applying P to (A8), (A10), and (A12) proves both claims for F_x. □

Proposition 1 (Monotonicity and capability option value)**.** *Assume $\widehat{g}_{t,i} \geq 0$ and the conditions of Lemma 1 at every period. Then the Bellman value $\widehat{V}_t(q, b)$ is nondecreasing in competence q and remaining capacity b. Moreover, for $b > 0$,*

$$\Omega_t(q, b) = \gamma[\widehat{V}_{t+1}(F_1(q), b-1) - \widehat{V}_{t+1}(F_0(q), b-1)] \geq 0. \tag{20}$$

Proof. Define the skip and review branches

$$S_t(q, b) := \gamma \widehat{V}_{t+1}(F_0(q), b), \tag{A13}$$
$$R_{t,i}(q, b) := q\widehat{g}_{t,i} - c_{t,i} + \gamma \widehat{V}_{t+1}(F_1(q), b-1). \tag{A14}$$

We use backward induction. At $t = T+1$, $\widehat{V}_{T+1}(q, b) = 0$ for every (q, b), hence it is nondecreasing in both arguments. Assume $\widehat{V}_{t+1}$ is nondecreasing in q. For $q' \geq q$,

$$S_t(q', b) - S_t(q, b) = \gamma[\widehat{V}_{t+1}(F_0(q'), b) - \widehat{V}_{t+1}(F_0(q), b)] \tag{A15}$$
$$\geq 0, \tag{A16}$$

where Lemma 1 gives $F_0(q') \geq F_0(q)$ and the induction hypothesis gives (A16). Likewise,

$$R_{t,i}(q', b) - R_{t,i}(q, b) = (q' - q)\widehat{g}_{t,i} \tag{A17}$$
$$+ \gamma[\widehat{V}_{t+1}(F_1(q'), b-1) - \widehat{V}_{t+1}(F_1(q), b-1)] \tag{A18}$$
$$\geq 0. \tag{A19}$$

The first summand is nonnegative because both factors are nonnegative; the second is nonnegative by transition and value monotonicity. For any two collections satisfying $f_j(q') \geq f_j(q)$ for every branch j, $\max_j f_j(q') \geq \max_j f_j(q)$. Applying this fact to S_t and all feasible $R_{t,i}$ proves competence monotonicity at t.

For capacity, let $\Pi_t(b)$ be the feasible continuation policies using at most b reviews. If $b' \geq b$, every $\pi \in \Pi_t(b)$ also belongs to $\Pi_t(b')$, hence $\Pi_t(b) \subseteq \Pi_t(b')$. Therefore

$$\widehat{V}_t(q, b') = \max_{\pi \in \Pi_t(b')} U_t(q; \pi) \tag{A20}$$
$$\geq \max_{\pi \in \Pi_t(b)} U_t(q; \pi) \tag{A21}$$
$$= \widehat{V}_t(q, b). \tag{A22}$$

Finally, Lemma 1 gives $F_1(q) \geq F_0(q)$. Applying the competence monotonicity just proved at $t+1$ yields

$$\widehat{V}_{t+1}(F_1(q), b-1) \geq \widehat{V}_{t+1}(F_0(q), b-1), \tag{A23}$$
$$\Omega_t(q, b) \geq 0 \tag{A24}$$

after subtracting the right side and multiplying by $\gamma \geq 0$. When $b = 0$, the review branch is infeasible. When $b > 0$, (A24) is only the value of retained capability; review is chosen only if some complete branch (A14) exceeds (A13). □

B Dynamic score decomposition, screening, and finite-horizon error

This section makes precise the interface between the paired-replay score and the dynamic controller. Candidate-local competence s_i may enter the score as a pre-intervention covariate. Separately, let

$q_0 > 0$ be the observed episode competence at which the complete candidate score is anchored for dynamic scheduling. Write its gross paired-replay protection as $P_{t,i}(q_0)$ and its observable review cost as $c_{t,i}$. In the paired-replay long-horizon benchmark, $P_{t,i}$ includes the candidate-local, pre-specified horizon interaction used to construct that replay. It is a single gross label at q_0 and is extended across competence states by $P_{t,i}(q) = (q/q_0)P_{t,i}(q_0)$. Rewards from later arriving workflow pools enter at their corresponding decision times. The net target is

$$B_{t,i} = P_{t,i}(q_0) - c_{t,i}, \qquad g_{t,i} := \frac{B_{t,i} + c_{t,i}}{q_0}. \tag{D1}$$

The implementation applies the same transformation to the learned score and uses $R_{t,i}(q) = qg_{t,i} - c_{t,i}$ in the Bellman recursion.

Proposition 2 (Reference-state recovery and disjoint accounting)**.** *The transformation in (D1) exactly recovers the paired-replay target at the reference state: $R_{t,i}(q_0) = B_{t,i}$. Moreover, suppose that (i) $B_{t,i}$ contains only outcomes assigned to the current candidate's replay horizon and (ii) rewards from pools arriving after t are assigned only to their own decision times. Then the discounted return*

$$J(\pi) = \sum_{t=1}^{T} \gamma^{t-1} \mathbf{1}\{a_t \neq 0\} R_{t,a_t}(q_t) \tag{D2}$$

counts every assigned replay reward once under this declared extension; the transition F changes the coefficient q_s of later rewards but contributes no immediate reward.

Proof. Substitution of (D1) into the reward at q_0 gives

$$\begin{aligned} R_{t,i}(q_0) &= q_0 g_{t,i} - c_{t,i} && \text{(D3)} \\ &= q_0 \frac{B_{t,i} + c_{t,i}}{q_0} - c_{t,i} && \text{(D4)} \\ &= B_{t,i} + c_{t,i} - c_{t,i} && \text{(D5)} \\ &= B_{t,i}. && \text{(D6)} \end{aligned}$$

The passage from (D3) to (D4) substitutes the definition of $g_{t,i}$; (D5) cancels the nonzero q_0; and (D6) cancels the cost added before normalization with the cost subtracted by the reward.

Fix a policy trajectory $\tau = (q_1, a_1, \ldots, q_T, a_T)$. By assumption (i), the outcome assigned at time t appears only in $R_{t,a_t}(q_t)$. By assumption (ii), every later-pool outcome appears only in some $R_{s,a_s}(q_s)$ with $s > t$. Expanding (D2) yields

$$\begin{aligned} J(\tau) = \mathbf{1}\{a_1 \neq 0\} R_{1,a_1}(q_1) + \gamma \mathbf{1}\{a_2 \neq 0\} R_{2,a_2}(q_2) + \cdots && \text{(D7)} \\ + \gamma^{T-1} \mathbf{1}\{a_T \neq 0\} R_{T,a_T}(q_T). && \text{(D8)} \end{aligned}$$

Each time index occurs exactly once. The recursion $q_{t+1} = F(q_t, \mathbf{1}\{a_t \neq 0\}, d_t)$ determines the argument of later reward functions but adds no summand to (D7)–(D8). Hence it changes future protection through competence, with each assigned replay reward counted once. The assumptions state the required accounting convention; a replay label that already included a realized reward from a later pool would require that term to be removed from one side of the interface. □

We next collect estimator, graph, transition, and competence-grid error in one finite-horizon statement. Let M denote the true model and $\widehat{M}_h$ the estimated model whose next state is projected to a grid with maximum nearest-neighbor error $h/2$. Costs are observed exactly. Assume uniformly over feasible (t, q, b, i) that

$$|g_{t,i} - \widehat{g}_{t,i}| \leq \epsilon_t^{\text{est}} + \epsilon_t^{\text{graph}} =: e_t^g, \tag{D20}$$

$$|F_x(q) - \widehat{F}_x(q)| \leq e_t^F, \qquad x \in \{0, 1\}. \tag{D21}$$

Here $\epsilon_t^{\text{graph}}$ may be instantiated by the omitted and spurious effect masses in Proposition 11 after the same competence normalization.

The bounded-error and Lipschitz conditions are stated uniformly over the reachable calibration envelope. They are natural here because q lies in the compact interval $[q_{\min}, 1]$, the clipped affine

transition is Lipschitz, the horizon and candidate set are finite, and gross gains are bounded by the replay utility scale. For any fixed action suffix, repeated composition and finite discounted summation therefore produce a finite continuation Lipschitz constant. In deployment, held-out residual and transition intervals supply e_t^g and e_t^F; the coverage gate in Appendix J.2 restricts use of the bound to this envelope. The sign margin below is likewise observable from the predicted reward and the calibrated error radius.

Proposition 3 (Unified finite-horizon policy regret)**.** *Suppose the continuation return of every feasible fixed action suffix is L_{t+1}-Lipschitz in competence. Both models use the same deterministic arrival forecast, initial state, and discount $\gamma \in [0,1]$. In addition, assume their positive-reward gates agree at every reachable candidate-state pair. A sufficient margin condition is*

$$|\widehat{R}_{t,i}(q)| > e_t^g \quad \textit{for every reachable } (t,i,q), \tag{D22}$$

because $|R_{t,i}(q) - \widehat{R}_{t,i}(q)| \le e_t^g$ then prevents a sign change. Define

$$\mathcal{E}_t := \sum_{s=t}^{T} \gamma^{s-t} \left[e_s^g + \gamma L_{s+1}(e_s^F + h/2) \right], \qquad L_{T+1} = 0. \tag{D23}$$

If $\widehat{\pi}$ is the action sequence returned by $\widehat{M}_h$ and π^* is the optimal action sequence in M, then

$$J_M^{\pi^*}(q_t, b) - J_M^{\widehat{\pi}}(q_t, b) \le 2\mathcal{E}_t. \tag{D24}$$

Proof. For deterministic transitions and an observed arrival forecast, each policy rollout induces a feasible action sequence. We first bound the value difference for an arbitrary sequence $\pi = (a_t, \dots, a_T)$. Let

$$D_t^\pi := \sup_{q,b} |J_{M,t}^\pi(q,b) - J_{\widehat{M}_h,t}^\pi(q,b)|. \tag{D25}$$

At $t = T+1$, both terminal values are zero, so $D_{T+1}^\pi = 0$. For the fixed action a_t, write $r_t, \widehat{r}_t$ for its rewards and $q', \widehat{q}'_h$ for its next states. The triangle inequality gives

$$\begin{aligned}
&|J_{M,t}^\pi(q,b) - J_{\widehat{M}_h,t}^\pi(q,b)| && \text{(D26)}\\
&= |r_t - \widehat{r}_t + \gamma[J_{M,t+1}^\pi(q',b') - J_{\widehat{M}_h,t+1}^\pi(\widehat{q}'_h,b')]| && \text{(D27)}\\
&\le |r_t - \widehat{r}_t| + \gamma |J_{M,t+1}^\pi(q',b') - J_{M,t+1}^\pi(\widehat{q}'_h,b')| && \text{(D28)}\\
&\quad + \gamma |J_{M,t+1}^\pi(\widehat{q}'_h,b') - J_{\widehat{M}_h,t+1}^\pi(\widehat{q}'_h,b')| && \text{(D29)}\\
&\le e_t^g + \gamma L_{t+1}|q' - \widehat{q}'_h| + \gamma D_{t+1}^\pi. && \text{(D30)}
\end{aligned}$$

For review, (D20) and $q \in [0,1]$ imply $|r_t - \widehat{r}_t| = q|g_{t,i} - \widehat{g}_{t,i}| \le e_t^g$; for skipping the difference is zero. If Π_h is nearest-grid projection, then

$$\begin{aligned}
|q' - \widehat{q}'_h| &= |F_x(q) - \Pi_h(\widehat{F}_x(q))| && \text{(D31)}\\
&\le |F_x(q) - \widehat{F}_x(q)| + |\widehat{F}_x(q) - \Pi_h(\widehat{F}_x(q))| && \text{(D32)}\\
&\le e_t^F + h/2. && \text{(D33)}
\end{aligned}$$

Taking the supremum in (D30), using (D33), and setting $d_t = e_t^g + \gamma L_{t+1}(e_t^F + h/2)$ yields

$$D_t^\pi \le d_t + \gamma D_{t+1}^\pi. \tag{D34}$$

Repeated substitution, followed by $D_{T+1}^\pi = 0$, gives

$$\begin{aligned}
D_t^\pi &\le d_t + \gamma d_{t+1} + \cdots + \gamma^{T-t} d_T && \text{(D35)}\\
&= \sum_{s=t}^{T} \gamma^{s-t}[e_s^g + \gamma L_{s+1}(e_s^F + h/2)] && \text{(D36)}\\
&= \mathcal{E}_t. && \text{(D37)}
\end{aligned}$$

Finally, add and subtract the estimated-model policy values:

$$\begin{aligned}
J_M^{\pi^*} - J_M^{\widehat{\pi}} &= [J_M^{\pi^*} - J_{\widehat{M}_h}^{\pi^*}] + [J_{\widehat{M}_h}^{\pi^*} - J_{\widehat{M}_h}^{\widehat{\pi}}] + [J_{\widehat{M}_h}^{\widehat{\pi}} - J_M^{\widehat{\pi}}] && \text{(D38)}\\
&\le |J_M^{\pi^*} - J_{\widehat{M}_h}^{\pi^*}| + 0 + |J_{\widehat{M}_h}^{\widehat{\pi}} - J_M^{\widehat{\pi}}| && \text{(D39)}\\
&\le \mathcal{E}_t + \mathcal{E}_t = 2\mathcal{E}_t. && \text{(D40)}
\end{aligned}$$

The middle term is nonpositive because $\widehat{\pi}$ is optimal among feasible action sequences in $\widehat{M}_h$; (D37) applies to each fixed sequence. □

C SENSITIVITY TO COMPETENCE OBSERVATION ERROR

The main recursion uses an observed competence state. The following bound separates error in that observation from error in predicted intervention gain.

Proposition 4 (Competence-error decision bound)**.** *Assume* $0 \le \widehat{g}_{t,i} \le G$*, and for* $x \in \{0,1\}$*,* $|F_x(q) - F_x(q')| \le L_F|q - q'|$*. Define*

$$L_t := G\sum_{j=0}^{T-t}(\gamma L_F)^j, \qquad L_{T+1} := 0. \tag{21}$$

Every Bellman action value and $\widehat{V}_t(\cdot, b)$ *is* L_t*-Lipschitz in* q*. If* $|\widetilde{q} - q| \le \epsilon$ *and* $\widetilde{a} \in \arg\max_a Q_t(\widetilde{q}, b, a)$*, then*

$$\max_a Q_t(q, b, a) - Q_t(q, b, \widetilde{a}) \le 2L_t\epsilon. \tag{22}$$

Proof. We first prove Lipschitz continuity by backward induction. At $t = T + 1$, $\widehat{V}_{T+1}(q, b) = 0$, hence

$$|\widehat{V}_{T+1}(q, b) - \widehat{V}_{T+1}(q', b)| = 0 = L_{T+1}|q - q'|. \tag{S1}$$

Assume $\widehat{V}_{t+1}$ is L_{t+1}-Lipschitz. For the skip branch $Q_t^0(q, b) = \gamma\widehat{V}_{t+1}(F_0(q), b)$,

$$\begin{aligned}
|Q_t^0(q, b) - Q_t^0(q', b)| &= \gamma|\widehat{V}_{t+1}(F_0(q), b) - \widehat{V}_{t+1}(F_0(q'), b)| && \text{(S2)}\\
&\le \gamma L_{t+1}|F_0(q) - F_0(q')| && \text{(S3)}\\
&\le \gamma L_{t+1}L_F|q - q'| && \text{(S4)}\\
&\le L_t|q - q'|. && \text{(S5)}
\end{aligned}$$

(S3) applies the induction hypothesis, (S4) transition Lipschitzness, and (S5) $L_t = G + \gamma L_F L_{t+1}$. For review candidate i, cancellation of the common cost followed by the triangle inequality gives

$$\begin{aligned}
|Q_t^i(q, b) - Q_t^i(q', b)| &= \big|(q - q')\widehat{g}_{t,i} + \gamma[\widehat{V}_{t+1}(F_1(q), b-1) - \widehat{V}_{t+1}(F_1(q'), b-1)]\big| && \text{(S6)}\\
&\le |q - q'|\,|\widehat{g}_{t,i}| + \gamma|\widehat{V}_{t+1}(F_1(q), b-1) - \widehat{V}_{t+1}(F_1(q'), b-1)| && \text{(S7)}\\
&\le [G + \gamma L_{t+1}L_F]|q - q'| && \text{(S8)}\\
&= L_t|q - q'|. && \text{(S9)}
\end{aligned}$$

For any finite family of L-Lipschitz functions,

$$\begin{aligned}
\max_a f_a(q) - \max_a f_a(q') &\le \max_a[f_a(q) - f_a(q')] && \text{(S10)}\\
&\le L|q - q'|. && \text{(S11)}
\end{aligned}$$

Swapping q, q' supplies the reverse inequality. Applying (S10)–(S11) to the skip and feasible review branches proves that $\widehat{V}_t$ is L_t-Lipschitz and closes the induction. Expanding $L_t = G + \gamma L_F L_{t+1}$ gives the stated sum.

For the decision bound, choose $a^* \in \arg\max_a Q_t(q, b, a)$ and add and subtract action values at $\widetilde{q}$:

$$\begin{aligned}
Q_t(q, b, a^*) - Q_t(q, b, \widetilde{a}) &= [Q_t(q, b, a^*) - Q_t(\widetilde{q}, b, a^*)] && \text{(S12)}\\
&\quad + [Q_t(\widetilde{q}, b, a^*) - Q_t(\widetilde{q}, b, \widetilde{a})] && \text{(S13)}\\
&\quad + [Q_t(\widetilde{q}, b, \widetilde{a}) - Q_t(q, b, \widetilde{a})]. && \text{(S14)}
\end{aligned}$$

The middle term is nonpositive because $\widetilde{a}$ maximizes at $\widetilde{q}$. Each remaining term is at most $L_t|q - \widetilde{q}| \le L_t\epsilon$ by the action-value result. Their sum is at most $2L_t\epsilon$. □

C.1 OBSERVED-STATE NOISE STRESS TEST

We complement the worst-case bound with a paired noise experiment. For each of the three estimator seeds and 100 shared 16-period episodes per source, the planner observes $\widetilde{q}_1 = \mathrm{clip}(q_1 + \sigma_q Z, .25, 1)$ with $Z \sim \mathcal{N}(0, 1)$, while the oracle, reward evaluation, and competence trajectory use the true q_1. The same standardized draw is scaled across noise levels within an episode. Table 5 reports

Table 5: Initial-competence observation noise at capacity four (300 episodes per source and noise level).

Source	σ_q	Regret	changed episode	action difference
A	.00	.0211	.000	.0000
A	.05	.0216	.053	.0067
A	.10	.0215	.063	.0079
B	.00	.0172	.000	.0000
B	.05	.0169	.047	.0063
B	.10	.0170	.063	.0083

mean regret, the fraction of episodes whose 16-action schedule differs anywhere from the zero-noise schedule, and the mean fraction of differing actions.

The perturbation changes relatively few actions because most tested Bellman comparisons have a decision margin larger than the induced value change. Small non-monotonic differences in mean regret fall within episode variation, so this experiment supports schedule stability under initial-state error. Receding-horizon error at every period defines a complementary operating condition.

D IDENTIFICATION OF THE PAIRED REPLAY ESTIMAND

For candidate i, let $J_i^A(\omega)$ and $J_i^{AH}(\omega)$ be the AUTO and REVIEW continuation losses under exogenous randomness ω. A replay pair restores the same pre-action state $X_i = (H_i, G_i, a_i, q_i, r_i)$, including the pre-decision reviewer descriptor r_i, and evaluates both branches with a common draw ω_r; repeated pairs use independent draws $\omega_1, \ldots, \omega_R$ conditional on X_i.

Proposition 5 (Unbiased paired-replay label)**.** *Assume consistency, common branch-independent exogenous randomness, no cross-workflow interference, and* $\mathbb{E}[|J_i^A| + |J_i^{AH}| \mid X_i] < \infty$. *Then*

$$\widehat{\Delta}_i = \frac{1}{R}\sum_{r=1}^{R}[J_i^A(\omega_r) - J_i^{AH}(\omega_r)] \tag{23}$$

satisfies $\mathbb{E}[\widehat{\Delta}_i \mid X_i] = \Delta_i(q_i)$.

Proof. Using linearity of conditional expectation,

$$\mathbb{E}[\widehat{\Delta}_i \mid X_i] = \mathbb{E}\left[\frac{1}{R}\sum_{r=1}^{R}(J_i^A(\omega_r) - J_i^{AH}(\omega_r)) \,\middle|\, X_i\right] \tag{I1}$$

$$= \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}[J_i^A(\omega_r) - J_i^{AH}(\omega_r) \mid X_i] \tag{I2}$$

$$= \frac{1}{R}\sum_{r=1}^{R}\Delta_i(q_i) \tag{I3}$$

$$= \Delta_i(q_i). \tag{I4}$$

(I2) uses integrability; (I3) uses consistency and the common conditional law of ω_r in both branches. Shared randomness is not required for unbiasedness if the marginal laws remain equal, but it reduces paired-label variance:

$$\mathrm{Var}(J_i^A - J_i^{AH} \mid X_i) = \mathrm{Var}(J_i^A \mid X_i) + \mathrm{Var}(J_i^{AH} \mid X_i) - 2\,\mathrm{Cov}(J_i^A, J_i^{AH} \mid X_i). \tag{24}$$

Positive covariance from common random numbers therefore removes shared environment variation. □

The proposition applies when restored states match, REVIEW follows the defined downstream policy, workflows are mutually noninterfering, and branch retention follows the pre-specified rule. Conditioning on r_i preserves intervention identity when outcomes span distinct reviewer policies. Within this design, $h_i^R - h_i^A$ serves as a predictor trained against $\widehat{\Delta}_i$; causal identification comes from the paired replay. The replay audit and exclusion rule in Section 4 verify these conditions.

Table 6: Seven-seed baseline stability (mean with sample SD in parentheses).

Method	MAE	ρ	Source A Regret@8	Source B Regret@8
HIVE	.0679 (.0109)	.8944 (.0195)	.0844 (.0568)	.0360 (.0330)
Twin-Outcome	.1150 (.0053)	.8894 (.0137)	.1712 (.0118)	.0715 (.0192)
Free-Delta	.0363 (.0084)	.8381 (.0452)	.1899 (.0432)	.1342 (.0467)
Direct-Policy	.0501 (.0054)	.8893 (.0138)	.2093 (.0298)	.1574 (.1811)
DeCCaF	.0586 (.0297)	.3543 (.4376)	.5316 (.3121)	.7003 (.4016)
FLORA-Marginal	.0431 (.0002)	.7555 (.1265)	.1739 (.0990)	.2034 (.0997)
L2D	.4700 (.0225)	.8418 (.0056)	.2772 (.0629)	.6492 (.0527)
CAIR-Mask	.0443 (.0012)	-.0666 (.2305)	.7030 (.2637)	1.0010 (.2511)

D.1 Alternative identification from a single logged branch

In a single-branch logging regime, each candidate reveals $Y = J^A$ when $R = 0$ or $Y = J^{AH}$ when $R = 1$. This is a distinct regime from the experiments above. Let $e(X) = \Pr(R = 1 \mid X)$ and $m_r(X) = \mathbb{E}[Y \mid R = r, X]$, where X comprises pre-review information. Under consistency, conditional exchangeability $(J^A, J^{AH}) \perp R \mid X$, and overlap $0 < e(X) < 1$, define

$$\phi(X, R, Y) = m_0(X) - m_1(X) - \frac{R}{e(X)}[Y - m_1(X)] + \frac{1 - R}{1 - e(X)}[Y - m_0(X)]. \tag{I5}$$

This uses the paper's protection-value sign convention, AUTO loss minus REVIEW loss. In practice, all nuisance functions must be cross-fitted so the label for an observation is produced by models that did not train on it (Chernozhukov et al., 2018).

Proposition 6 (Doubly robust logged-branch score)**.** *Let e^*, m_0^*, m_1^* denote the true nuisance functions. Conditional on X, $\mathbb{E}[\phi \mid X] = \mathbb{E}[J^A - J^{AH} \mid X]$ if either (i) $e = e^*$, or (ii) $m_0 = m_0^*$ and $m_1 = m_1^*$.*

Proof. By iterated expectation and $\mathbb{E}[RY \mid X] = e^* m_1^*$,

$$\mathbb{E}\left[\frac{R}{e}(Y - m_1)\middle| X\right] = \frac{e^*}{e}(m_1^* - m_1), \tag{I6}$$

$$\mathbb{E}\left[\frac{1 - R}{1 - e}(Y - m_0)\middle| X\right] = \frac{1 - e^*}{1 - e}(m_0^* - m_0). \tag{I7}$$

Substitution into (I5) gives

$$\mathbb{E}[\phi \mid X] = m_0 - m_1 - \frac{e^*}{e}(m_1^* - m_1) + \frac{1 - e^*}{1 - e}(m_0^* - m_0). \tag{I8}$$

If $e = e^*$, the two ratios reduce to one, so (I8) equals $m_0^* - m_1^*$. If both outcome regressions are correct, both residual differences in (I8) vanish and the same equality follows. Exchangeability and consistency then give $m_0^* - m_1^* = \mathbb{E}[J^A - J^{AH} \mid X]$. □

This route applies under overlap, recorded assignment factors, and a stable REVIEW intervention across operators. Propensity histograms, effective sample size, and outcome-model residuals operationalize these requirements in the applicability gate below before logged labels replace paired replay.

Part II: Experimental protocol and additional evidence

E Additional static, shift, and one-step results

Each learned baseline and HIVE use seeds 11, 17, 23, 29, 31, 37 and 41 in the headline static comparison. The resulting baseline means and standard deviations at capacity 8 are reported below; the main table reports the same statistics compactly.

E.1 CAPACITY-MATCHED UNCONSTRAINED DELTA-GNN

We retain the capacity-matched Free-Delta-GNN as an architectural ablation rather than a headline baseline. It matches HIVE's factual and delta GNN capacity, inputs, losses, and schedule, while freeing delta offsets, descendant support, and readout centering. Table 7 reports seven-seed means. Null violation is the mean absolute prediction after replacing REVIEW with AUTO, for which an intervention-attributable score should be zero.

Table 7: Strong matched-GNN ablation on ID and structural-shift tests.

Split	Method	Source A Regret@8	Source B Regret@8	Null violation
ID	HIVE	.0844	.0360	.0000
	Free-Delta-GNN	.0700	.0469	.0667
Shift	HIVE	.0312	.0424	.0000
	Free-Delta-GNN	.0585	.0442	.0742

On shifted Source A, HIVE has lower regret in 6/7 paired seeds (one-sided Wilcoxon $p = .039$). Source B is comparable under this shift. The exact null score and shifted Source-A advantage isolate the value of HIVE's structural constraints even though the unconstrained model is competitive on the ID pool.

The shift result shows a modest increase in pointwise error while preserving low allocation regret. Relative to training, the set jointly changes graph depth, reduces review detectability, and replaces the utility equation with sublinear risk and delayed exposure saturation. It therefore tests the frozen score when both review response and downstream value change.

E.2 CLAIM–EVIDENCE MAP

The paper ties each claim to a distinct check. Intervention attribution is an architectural statement proved by the null and descendant invariants; semantic usefulness is tested by mode removal, semantic shuffling, and the shifted response rule. Allocation quality is measured at the operational capacity, while intertemporal value is tested against one-step and no-retention comparisons.

Claim	Direct evidence or falsification check
Attributable intervention score	Null/support proposition; mode-removal and candidate-local ablations
Scarce-capacity ordering	Regret@K on shared pools; three-seed Twin-Outcome comparison
Changed-response transfer	Held-out depth/detectability/utility shift; shuffled REVIEW semantics
Intertemporal planning	HIVE-DP versus one-step/no-horizon; retention and arrival comparisons
Competence sensitivity	Transition grid, specialization test, and state-error bound

E.3 FROZEN REPRESENTATION SWAP

We replace the deterministic 128-dimensional unigram/bigram map by frozen 384-dimensional all-MiniLM-L6-v2 sentence representations (Wang et al., 2020). Text templates, graphs, labels, splits, HIVE layers, loss weights, and seeds remain fixed; the input representation and its projection dimension change. Table 9 reports three-seed means (standard deviations for regret).

The learned encoder preserves useful allocation ordering, including low Source B regret, while the deterministic hash representation gives lower Source-A variation. This result demonstrates compatibility across the two representation families. Independently authored review policies provide the direct test of semantic fidelity.

E.4 PLUG-IN POLICY-LEARNING BASELINE

Direct-Policy regresses the paired net effect from candidate-local AUTO/REVIEW representations and numeric features, then uses the same ranking and relaxed capacity losses as HIVE. This implements the plug-in empirical-welfare pattern (Kitagawa & Tetenov, 2018; Athey & Wager, 2021) at the candidate-local level. Across seeds 11/17/23/29/31/37/41 it obtains MAE .0501 (.0054), Spearman

Table 8: Review-response/depth/utility shift, evaluated with the seed-11 in-distribution checkpoint.

Source	MAE	ρ	Regret@4	Regret@8
Source A	.0933	.8734	.0119	.0244
Source B	.1005	.8727	.0207	.0291

Table 9: Frozen learned-representation swap. ID denotes the original test split; Shift changes response, depth, and utility.

Representation	Split	MAE	ρ	Source A Regret@8	Source B Regret@8
Hash	ID	.0663	.8963	.1244	.0147
MiniLM	ID	.0791	.8560	.1417 (.0648)	.0244 (.0250)
MiniLM	Shift	.0974	.8310	.0403 (.0151)	.0545 (.0159)

Table 10: Practice-dependent competence experiment, mean (standard deviation) over seeds 11, 17 and 23.

Configuration	MAE	ρ	Source A Regret@8	Source B Regret@8
HIVE-S	.0759 (.0082)	.8911 (.0082)	.0543 (.0276)	.0474 (.0444)
no competence state	.0649 (.0138)	.8734 (.0175)	.0648 (.0334)	.0837 (.0194)

$\rho = .8893$ (.0138), and Regret@8 .2093 (.0298)/.1574 (.1811) on Sources A/B. Its good pointwise and rank metrics yield higher allocation regret than HIVE, isolating the contribution of intervention structure beyond shared decision-focused supervision.

HIVE-S reduces mean regret relative to the parameter-matched no-competence ablation by 16.2% on Source A and 43.4% on Source B. Its MAE is higher, again separating accurate value averaging from correct scarce-capacity ordering. This experiment validates the value of an observed competence state under the pre-specified skill dynamics. The coefficients and transition are reported in full so they can be replaced by longitudinal estimates in a deployment study.

E.5 Long-horizon competence retention

The static and practice-aware experiments treat each pool independently. We next test the temporal increment in Section 3: a review can also preserve capability that will be needed for a forecasted future workload. For each candidate we expose a normalized review horizon h_t and future review pressure p_t before allocation. The pre-specified target adds the discounted retention term from Eq. 9, with the fixed interaction $(1 + 1.5h_t p_t)$; the resulting protocol gives a reproducible long-horizon stress test for the proposed state variables.

HIVE-L uses the eight numeric features (the four node features, practice frequency, competence, horizon, and pressure) and applies the pre-specified retention gate to the counterfactual delta. The parameter-matched *no-horizon* model retains the same architecture and competence features but masks h_t, p_t at both training and test time. We report the mean and standard deviation over three seeds; the allocation pool, capacity, and task-cluster bootstrap are unchanged from the main experiment.

Forecast conditioning alone has a modest source-dependent effect in the one-step allocator. This motivates evaluating the setting for which the state was introduced: carrying competence and remaining capacity across an arrival sequence.

F Complete experimental specification

This section records the configuration needed to reproduce the paired-replay suite results. The long-horizon manifest uses seed 20260829. There are six task families per source, 20 instances per family, and 240 workflows in total (120 per source). Splitting is by task identity: 60%/20%/20% for train/validation/test, giving 144/48/48 workflows and 720 candidate interventions. Each training candidate has one potential-outcome repetition; validation and test candidates have three noisy repetitions whose mean is the target. The shift set contains 160 held-out workflows with deeper traces, lower detectability, and a changed utility equation.

Table 11: Long-horizon retention experiment (three seeds; capacity 8 regret, lower is better).

Configuration	MAE	ρ	Source A Regret@8	Source B Regret@8
HIVE-L	.0690 (.0071)	.8478 (.0606)	.0597 (.0210)	**.0484** (.0111)
no horizon state	.0699 (.0101)	.8771 (.0061)	**.0418** (.0288)	.0515 (.0146)

The representation is fixed across methods. Each source contributes six task families and 120 workflows; traces contain 8–12 nodes in the long-horizon split. Deterministic unigram/bigram hashing produces 128-dimensional mode representations. Every candidate additionally has eight numeric inputs: four node features plus practice, competence, remaining-horizon, and review-pressure features. The graph distinguishes control, state, and entity relations. The three HIVE checkpoints use seeds 11, 17, and 23; all 720 candidates retain the same split identity across seeds.

For completeness, the benchmark constructor samples risk levels $r \in \{0, 1, 2, 3\}$ with probabilities $(.42, .31, .19, .08)$ for read-like nodes and $(.28, .32, .25, .15)$ for state-changing nodes. Their risk weights are $(.08, .28, .58, .88)$ and review detectabilities are $(.05, .34, .70, .91)$. The propagation exposure used to construct paired labels is computed by a forward recurrence. Set $z_i = 1$ and $z_v = 0$ for other nodes, process nodes in index order, and update each outgoing edge $e : u \to v$ by

$$z_v \mathrel{+}= \min(1.5, z_u) \frac{\alpha_{\mathrm{type}(e)}}{\sqrt{\max(1, \deg^-(v))}}, \qquad E_i = \sum_{v \succeq i} z_v w_{\mathrm{type}(v)}. \tag{25}$$

Edge weights are $(.10, .72, .42)$ for control/state/entity edges; node criticalities are $(.15, .45, .85, 1.00, 1.20, .25)$ for READ/COMPUTE/WRITE/COMMUNICATE/SUBMIT/OTHER.

The pre-review practice frequency is sampled from $\mathrm{Beta}(2.4, 2.0)$, and competence is

$$q = \max\{.25, \min\{1, .25 + .75 \exp[-1.8(1 - f)]\}\}. \tag{26}$$

For long-horizon records, $h \sim \mathrm{Uniform}(.25, 1)$, $p \sim \mathrm{Beta}(2, 2)$, and the discount is $\exp[-.35(1 - h)]$. The non-shift causal and retention components are

$$g_i^{\mathrm{causal}} = q_i r_i \min(.96, d_i)(.055 + .060 E_i), \tag{27}$$

$$g_i^{\mathrm{ret}} = .05(1 - q_i) \exp[-.35(1 - h_i)](.5 + .5 d_i)(1 + 1.5 p_i h_i). \tag{28}$$

The stored paired value is their sum minus review overhead (.026, plus .018 for level-0 cases), plus Gaussian noise with standard deviation .018, clipped to $[-.075, .44]$. The public target is reconstructed from $J = 1 - S + .05 N_{\mathrm{step}}/N_{\max} + .02 C_{\mathrm{review}}$; model and planner inputs comprise the declared pre-intervention information set.

F.1 Five-source structural expansion

We separately test whether the static estimator survives a larger and more heterogeneous environment design. This expansion has 1,500 workflows and 4,500 candidate interventions: 300 workflows, 900 candidates, and six task families for each of five sources. Identity-based 60/20/20 splitting gives 900/300/300 workflows for training/validation/test, including 60 test workflows per source. The source labels denote interaction contracts distilled before model training from public task specifications. Each contract retains the operational data flow, action preconditions, and external-effect pattern that distinguishes its source; the constructor varies instances within this fixed task requirement rather than selecting structures from model behavior.

The first two contracts extend the company-operation and policy/service requirements used in the main two-source suite. Cross-app structure retains the multi-application state joins exercised by AppWorld (Trivedi et al., 2024); policy dialogue retains the user–policy–tool separation in τ-bench (Yao et al., 2024); and stateful tools retain the prerequisite, canonicalization, insufficient-information, and distractor categories of ToolSandbox (Lu et al., 2025). BrowserGym's common interface across web environments motivates a shared observation/action schema while the source tasks determine the dependency graph (Le Sellier De Chezelles et al., 2025). Every graph is instantiated from its declared task contract, so source-specific dependencies precede and remain independent of estimator fitting.

For candidate i, let d_i be the number of reachable descendants, e_i^S, e_i^E the counts of reachable state and entity edges, q reviewer competence, r_i local risk, and p_i evidence detectability. The additional

Table 12: Source-specific structures in the expanded suite. Edge counts vary across instances because long-range dependencies are sampled.

Source contract	Distinguishing structure	$\|V\|$	$\|E\|$ range
Company DAG	branches, shared entities, and downstream state reuse	9	15–23
Service gate	policy lookup followed by approval-gated mutation	9	13–21
Cross-app	two-app query, record join, fan-out writes, and fan-in check	10	21–30
Policy dialogue	user clarification, policy constraint, irreversible tool call	9	17–25
Stateful tools	latent prerequisite, canonicalization, milestone, distractor branch	10	17–27

binary cues are policy ℓ_i, approval gate a_i, cross-app join x_i, ambiguity u_i, irreversible mutation z_i, missing prerequisite m_i, canonicalization c_i, and distractor branch h_i. Before observation noise and source-specific overhead, the five protection responses are

$$G_i^{\text{company}} = qr_i p_i[.030 + .020d_i + .050\min(4, e_i^S) + .025\min(3, e_i^E)], \tag{29}$$

$$G_i^{\text{service}} = qr_i p_i[.025 + .016d_i + .040\min(4, e_i^S) + .055\ell_i + .045a_i], \tag{30}$$

$$G_i^{\text{cross}} = qr_i^{.75} p_i[.025 + .016d_i + .050\min(4, e_i^E) + .050x_i], \tag{31}$$

$$G_i^{\text{dialogue}} = qr_i p_i[.025 + .015d_i + .040\min(4, e_i^S) + .045u_i + .055\ell_i + .040z_i], \tag{32}$$

$$G_i^{\text{stateful}} = qr_i p_i[.025 + .016d_i + .040\min(4, e_i^S) + .055m_i + .045c_i](1 - .22h_i). \tag{33}$$

The outcome is $B_i = \text{clip}(G_i - o_i + \epsilon_i, -.09, .42)$, $\epsilon_i \sim \mathcal{N}(0, .02^2)$; o_i is fully specified in the released benchmark construction and ranges from .032 to .050 according to the interaction contract. Validation and test targets average three independent outcome draws. Each base contract also receives sampled long-range state/entity edges; these are visible to graph methods and vary within candidate positions. Consequently, the same apparent risk has different value under an approval gate, cross-app join, or missing state prerequisite, and its exposure varies across instances of the same family.

This extension audits scaling, schema coverage, and source-specific graph construction. The matched two-source study provides the confirmatory method comparison, while the expansion serves as a pre-specified structural audit.

F.2 Training and allocation configuration

All learned models use AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, gradient norm clipped at 1, 12 maximum epochs, and early stopping patience 3. The capacity-aligned loss uses decision capacity 8 and temperature .10.

The encoder has hidden dimension 64, three graph layers, dropout .1, and batches of 32 candidates drawn from distinct workflows. Value, ranking, and capacity losses receive weights $1/.25/.5$. Static evaluation samples pools of up to 32 workflows at capacities 4, 8, and 12. A source-specific test split has 24 workflows, so the cap selects the same complete source pool: nominal repeats are computational duplicates; uncertainty intervals therefore use 200 task-cluster bootstrap draws. Dynamic evaluation uses 16 periods, capacities 4 and 6, a 151-point competence grid, discount .97, and nominal transition parameters idle .05, decay 1.8, and retention .08. These quantities are read from the released configuration and are fixed before test evaluation.

The allocator receives the factual prefix, both frozen mode representations, the typed graph, numeric features, and observed initial competence. Its information set ends before the paired outcome and post-review trace. Static allocation solves the one-review-per-workflow integer problem exactly. Dynamic allocation uses Eq. 13; the online variant replans after each period using the current pool and arrival forecast. The uncertainty ensemble uses checkpoints from seeds 11, 17, and 23. Validation compares neutral, nominal lower-confidence, and worst-case-transition lower-confidence planners. The one-standard-error rule is applied per source; for a new source with too few validation cases, the pooled validation choice is the deployment fallback.

For a distinct-workflow batch $\mathcal{B}$, the HIVE objective is

$$\mathcal{L} = \lambda_v \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (\widehat{B}_i - B_i)^2 + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_d \mathcal{L}_{\text{cap}}, \tag{34}$$

with $(\lambda_v, \lambda_r, \lambda_d) = (1, .25, .5)$. For every ordered pair with $B_i > B_j$, the ranking term is $[.1 - (\widehat{B}_i - \widehat{B}_j)]_+$; pairs with equal targets are omitted. The capacity term is

$$\mathcal{L}_{\text{cap}} = -\frac{1}{K} \sum_i a_i B_i, \qquad a_i = \sigma((\widehat{B}_i - \tau)/.10), \tag{35}$$

where τ is solved so that the scores and K zero-valued dummy items have total relaxed mass K. The dummy items implement the at-most-K outside option and prevent the loss from rewarding negative-value reviews.

Baseline training uses the same optimizer, data split, early-stopping rule, and candidate pools as HIVE. All randomization is seeded; pool orderings are shared within a comparison and the workflow is the bootstrap unit. This makes the reported differences attributable to the estimator or scheduler rather than to different sampling or evaluation procedures.

F.3 Paired seed analysis for the static headline

We pair capacity-eight regret by the same seven training seeds. Positive differences favor HIVE. We report the mean paired improvement, HIVE's strict win count, and a two-sided paired t-test; the seed is the replication unit, with candidate nodes grouped within their workflow.

Table 13: Seed-paired static allocation analysis.

Comparator	Source	Mean improvement	Wins/7	p
Free-Delta	A	.1055	6	.0088
	B	.0982	7	.0083
Twin-Outcome	A	.0869	6	.0055
	B	.0355	6	.0568
Direct-Policy	A	.1249	7	.0021
	B	.1214	6	.1536

The tests support a stable advantage over the candidate-local Free-Delta on both sources and over Twin-Outcome and Direct-Policy on Source A. Source-B differences for the latter two controls are directionally favorable with wider seed variation. Appendix E separately evaluates the strong capacity-matched Free-Delta-GNN as an architectural ablation.

G Proof details and allocation-error guarantees

G.1 Transition monotonicity and Bellman induction

Lemma 1 and Proposition 1 give the complete transition and Bellman proofs, including clipping, the $b = 0$ boundary, and every induction step. The robust-transition and allocation results below build directly on these arguments.

G.2 Rectangular transition uncertainty

Proposition 7 (Worst transition in a rectangular set)**.** *Let $\Theta = [\rho_-, \rho_+] \times [\eta_-, \eta_+]$, and suppose the clipping interval is fixed. If $\widehat{V}_{t+1}$ is nondecreasing in competence, then for either action $x \in \{0, 1\}$,*

$$\min_{(\rho,\eta)\in\Theta} \widehat{V}_{t+1}(F_{\rho,\eta}(q, x, d), b - x) = \widehat{V}_{t+1}(F_{\rho_+,\eta_-}(q, x, d), b - x). \tag{36}$$

Proof. For the unclipped transition $\bar{F}_{\rho,\eta} = q_{\min} + (q - q_{\min})e^{-\rho d} + \eta x(1-q)$, direct differentiation gives

$$\frac{\partial \bar{F}_{\rho,\eta}}{\partial \rho} = (q - q_{\min})(-d)e^{-\rho d} \le 0, \tag{A25}$$

$$\frac{\partial \bar{F}_{\rho,\eta}}{\partial \eta} = x(1-q) \ge 0. \tag{A26}$$

Thus, for every $(\rho, \eta) \in \Theta$,

$$\bar{F}_{\rho_+,\eta_-}(q,x,d) \le \bar{F}_{\rho,\eta_-}(q,x,d) \tag{A27}$$

$$\le \bar{F}_{\rho,\eta}(q,x,d). \tag{A28}$$

The first inequality increases decay from ρ to ρ_+; the second increases retention from η_- to η. Applying the nondecreasing projection P preserves (A27)–(A28), and applying nondecreasing $\widehat{V}_{t+1}(\cdot, b-x)$ preserves them again. Hence the corner (ρ_+, η_-) is no larger than every feasible alternative and attains the minimum. Rectangularity is required so that this corner belongs to Θ at each period. In the experiment it equals $(2.2, .04)$. □

G.3 From score error to allocation regret

Proposition 8 (Static allocation regret)**.** *Let $\mathcal{F}_K$ be the feasible family of review sets. Let $U(S) = \sum_{i\in S} B_i$, $\widehat{U}(S) = \sum_{i\in S} \widehat{B}_i$, and assume $|B_i - \widehat{B}_i| \le \epsilon_i$. If $S^* \in \arg\max_{S\in\mathcal{F}_K} U(S)$ and $\widehat{S} \in \arg\max_{S\in\mathcal{F}_K} \widehat{U}(S)$, then*

$$U(S^*) - U(\widehat{S}) \le 2 \max_{S\in\mathcal{F}_K} \sum_{i\in S} \epsilon_i. \tag{37}$$

Proof. Add and subtract the two estimated utilities:

$$U(S^*) - U(\widehat{S}) = [U(S^*) - \widehat{U}(S^*)] + [\widehat{U}(S^*) - \widehat{U}(\widehat{S})] + [\widehat{U}(\widehat{S}) - U(\widehat{S})] \tag{A29}$$

$$\le [U(S^*) - \widehat{U}(S^*)] + [\widehat{U}(\widehat{S}) - U(\widehat{S})] \tag{A30}$$

$$\le |U(S^*) - \widehat{U}(S^*)| + |\widehat{U}(\widehat{S}) - U(\widehat{S})| \tag{A31}$$

$$= \left|\sum_{i\in S^*} (B_i - \widehat{B}_i)\right| + \left|\sum_{i\in\widehat{S}} (\widehat{B}_i - B_i)\right| \tag{A32}$$

$$\le \sum_{i\in S^*} |B_i - \widehat{B}_i| + \sum_{i\in\widehat{S}} |B_i - \widehat{B}_i| \tag{A33}$$

$$\le \sum_{i\in S^*} \epsilon_i + \sum_{i\in\widehat{S}} \epsilon_i \tag{A34}$$

$$\le 2 \max_{S\in\mathcal{F}_K} \sum_{i\in S} \epsilon_i. \tag{A35}$$

(A30) uses optimality of $\widehat{S}$, so $\widehat{U}(S^*) - \widehat{U}(\widehat{S}) \le 0$; (A31) uses $a + b \le |a| + |b|$; (A33) is the triangle inequality; (A34) applies the coordinate-wise error bounds; and (A35) bounds both feasible sets by the same maximum.

For dynamic rewards, $R_i(q) = qg_i - c_i$ and $q \in [0,1]$, so a gain error $|g_i - \widehat{g}_i|$ satisfies

$$|R_i(q) - \widehat{R}_i(q)| = |qg_i - c_i - (q\widehat{g}_i - c_i)| \tag{A36}$$

$$= q|g_i - \widehat{g}_i| \tag{A37}$$

$$\le |g_i - \widehat{g}_i|. \tag{A38}$$

Thus $\epsilon_i = \kappa\sigma_i$ controls immediate reward error when the ensemble interval is valid; transition error is handled by the preceding proposition. □

G.4 INTERVENTION-ATTRIBUTION INVARIANTS

Proposition 9 (Null intervention and descendant support)**.** *For candidate i, HIVE satisfies: (i) if $h_i^R = h_i^A$, then $\widehat{B}_i = 0$ for every factual graph and pre-review numeric feature vector; and (ii) for every layer $\ell \geq 0$, $d_v^{(\ell)} = 0$ whenever $v \not\succeq i$.*

Proof. For (ii), the initialization gives $d_v^{(0)} = 0$ for $v \neq i$, and i is its own reachable descendant. Assume the support statement holds at layer ℓ. Equation 7 multiplies every updated state by $\mathbf{1}[v \succeq i]$; hence for $v \not\succeq i$,

$$d_v^{(\ell+1)} = 0 \cdot \mathrm{LN}_0(\cdots) = 0. \tag{A39}$$

Induction proves descendant support at every layer, independently of the edge messages inside the parentheses.

For (i), from Eq. 5 and the bias-free input map L_0,

$$d_i^{(0)} = L_0(h_i^R - h_i^A) = L_0(0) = 0, \tag{A40}$$

$$d_v^{(0)} = 0 \quad (v \neq i). \tag{A41}$$

We prove $d_v^{(\ell)} = 0$ for all v by induction on layer ℓ. Equations (A40)–(A41) establish the base case. Assume $d_u^{(\ell)} = 0$ for every u. For each edge (u, v) of relation r,

$$W_{\ell,r} d_u^{(\ell)} \odot g_\ell(z_u, z_v) = W_{\ell,r} 0 \odot g_\ell(z_u, z_v) = 0. \tag{A42}$$

Summing and degree-normalizing zero messages gives $m_v^{(\ell)} = 0$. The self-map is bias-free, so $D_\ell d_v^{(\ell)} = D_\ell 0 = 0$; GELU maps zero to zero; and the non-affine normalization satisfies $\mathrm{LN}_0(0) = 0$. Consequently,

$$d_v^{(\ell+1)} = \mathbf{1}[v \succeq i]\mathrm{LN}_0(0 + \mathrm{GELU}(0 + 0)) \tag{A43}$$

$$= \mathbf{1}[v \succeq i]\mathrm{LN}_0(0) = 0. \tag{A44}$$

This closes the induction. Substituting $d_v^{(L)} = 0$ into Eq. 8 yields

$$\widehat{B}_i = \sum_{v \succeq i} [\rho(z_v, 0) - \rho(z_v, 0)] + [\beta(z_i, 0, n_i) - \beta(z_i, 0, n_i)] \tag{A45}$$

$$= \sum_{v \succeq i} 0 + 0 = 0. \tag{A46}$$

Thus neither factual difficulty, graph size, nor numeric features can create intervention value when the two mode representations coincide. □

Proposition 10 (Strict containment in unrestricted outcome differencing)**.** *Let $\mathcal{H}_{\mathrm{HIVE}}$ be HIVE's score class and let $\mathcal{H}_{\mathrm{T}} = \{f_A(G, X) - f_R(G, X, m)\}$ contain unrestricted factual and review outcome heads with additive biases. If the unrestricted review head can represent every HIVE score, then $\mathcal{H}_{\mathrm{HIVE}} \subsetneq \mathcal{H}_{\mathrm{T}}$.*

Proof. Take any $h \in \mathcal{H}_{\mathrm{HIVE}}$. By the stated expressivity assumption, the review head can realize $f_R = -h$; choosing $f_A = 0$ gives $f_A - f_R = h$. Since h was arbitrary, $\mathcal{H}_{\mathrm{HIVE}} \subseteq \mathcal{H}_{\mathrm{T}}$.

It remains to show that the inclusion is proper. The preceding proposition proves both structural restrictions for every member of $\mathcal{H}_{\mathrm{HIVE}}$. In $\mathcal{H}_{\mathrm{T}}$, choose $f_A(G, X) = 0$ and $f_R(G, X, m) = b$ for any constant $b \neq 0$. Their difference is $-b$ even when $m = 0$, violating null intervention. Alternatively, let f_R depend on a feature of a disconnected node while $f_A = 0$; the difference then violates descendant support. Hence $\mathcal{H}_{\mathrm{T}}$ contains functions excluded by $\mathcal{H}_{\mathrm{HIVE}}$. Combining this witness with the inclusion above proves $\mathcal{H}_{\mathrm{HIVE}} \subsetneq \mathcal{H}_{\mathrm{T}}$. This is a structural restriction, not an accuracy ordering between the classes. □

G.5 GRAPH MISSPECIFICATION AND ALLOCATION REGRET

Let $D_i^\star$ be the true affected set for candidate i, let D_i be the descendant set in the recorded graph, and write the true and estimated node contributions as ϕ_{iv} and $\widehat{\phi}_{iv}$. Thus $B_i = \sum_{v \in D_i^\star} \phi_{iv}$ and $\widehat{B}_i = \sum_{v \in D_i} \widehat{\phi}_{iv}$.

Proposition 11 (Graph-error decomposition and top-K regret)**.** *Define*

$$E_i = \left| \sum_{v \in D_i \cap D_i^\star} (\widehat{\phi}_{iv} - \phi_{iv}) \right|, \tag{G1}$$

$$O_i = \sum_{v \in D_i^\star \setminus D_i} |\phi_{iv}|, \qquad S_i = \sum_{v \in D_i \setminus D_i^\star} |\widehat{\phi}_{iv}|. \tag{G2}$$

Then $|\widehat{B}_i - B_i| \le E_i + O_i + S_i$. If this sum is at most δ for every candidate, selecting the feasible top-K set by $\widehat{B}_i$ incurs true allocation regret at most $2K\delta$.

Proof. Partitioning both supports gives

$$\widehat{B}_i - B_i = \sum_{v \in D_i \cap D_i^\star} (\widehat{\phi}_{iv} - \phi_{iv}) + \sum_{v \in D_i \setminus D_i^\star} \widehat{\phi}_{iv} - \sum_{v \in D_i^\star \setminus D_i} \phi_{iv}. \tag{G3}$$

Apply $|a+b+c| \le |a| + |b| + |c|$, followed by $|\sum_j z_j| \le \sum_j |z_j|$, to obtain

$$|\widehat{B}_i - B_i| \le E_i + \sum_{v \in D_i \setminus D_i^\star} |\widehat{\phi}_{iv}| + \sum_{v \in D_i^\star \setminus D_i} |\phi_{iv}| \tag{G4}$$

$$= E_i + S_i + O_i. \tag{G5}$$

Let $A^\star$ maximize the true sum and $\widehat{A}$ maximize the estimated sum over the same feasible family; each contains at most K candidates. Then

$$\sum_{i \in A^\star} B_i - \sum_{i \in \widehat{A}} B_i = \sum_{i \in A^\star} (B_i - \widehat{B}_i) + \left(\sum_{i \in A^\star} \widehat{B}_i - \sum_{i \in \widehat{A}} \widehat{B}_i \right) + \sum_{i \in \widehat{A}} (\widehat{B}_i - B_i). \tag{G6}$$

The middle term is nonpositive by the definition of $\widehat{A}$; (G5) bounds each remaining term by $K\delta$. Hence

$$\sum_{i \in A^\star} B_i - \sum_{i \in \widehat{A}} B_i \le K\delta + 0 + K\delta = 2K\delta. \tag{G7}$$

The same argument covers the paper's partition constraint because it only reduces the shared feasible family. □

This decomposition supports three diagnostics: edge deletion probes O_i, edge insertion probes S_i, and held-out contribution error probes E_i. The applicability gate in Appendix J.2 admits graphs whose omitted and spurious mass can be bounded within the declared tolerance.

One implementable audit uses a randomly sampled, double-annotated shadow set. Two domain annotators independently reconstruct dependencies while blinded to the production graph; disagreements are adjudicated, and the union of production and adjudicated edges defines an audit supergraph. Paired replay is then rerun after adding each adjudicated missing edge and after deleting each production-only edge. Summed absolute changes in continuation loss give sample analogues of O_i and S_i; their cluster-bootstrap upper confidence limits replace the population quantities in (G5). The resulting release gate requires adequate inter-annotator agreement, stabilized new-edge discovery, and upper limits within the declared allocation tolerance. Sources outside this envelope receive additional graph collection before model allocation.

The next result states the population certificate induced by a finite shadow set and its explicit coverage term. Sample workflows are the independent units, with individual nodes grouped within workflows. For workflow j, let Z_j be its chosen aggregate omitted-plus-spurious mass (for example,

the maximum over its candidate nodes), and assume $0 \le Z_j \le R$. The i.i.d. condition is imposed on sampled workflow clusters: all correlated nodes, edges, and candidate audits within one workflow are first reduced to the single statistic Z_j. It therefore permits arbitrary within-workflow dependence and is supported operationally by sampling task identities independently from the declared deployment population; repeated runs of the same task identity remain in one cluster.

Proposition 12 (Population transfer of a shadow-graph audit)**.** *Let $Z_1, \dots, Z_n$ be i.i.d. workflow-cluster audit values with mean μ and empirical mean $\bar{Z}_n$. For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$\mu \le \bar{Z}_n + R\sqrt{\frac{\log(1/\alpha)}{2n}}. \tag{G8}$$

Let $\widehat{F}_n$ be the empirical CDF and set $\varepsilon_n = \sqrt{\log(2/\alpha)/(2n)}$. On an event of probability at least $1-\alpha$, every z satisfies $F(z) \ge \widehat{F}_n(z) - \varepsilon_n$. Consequently, for $\tau > \varepsilon_n$ and

$$\widehat{q}_{1-\tau+\varepsilon_n} := \inf\{z : \widehat{F}_n(z) \ge 1 - \tau + \varepsilon_n\}, \tag{G9}$$

the population tail obeys $\Pr[Z > \widehat{q}_{1-\tau+\varepsilon_n}] \le \tau$.

Proof. Because $Z_j \in [0, R]$, Hoeffding's inequality gives, for every $u > 0$,

$$\Pr(\mu - \bar{Z}_n \ge u) \le \exp\left(-\frac{2nu^2}{R^2}\right). \tag{G10}$$

Choose $u = R\sqrt{\log(1/\alpha)/(2n)}$. Then

$$\exp\left(-\frac{2nu^2}{R^2}\right) = \exp\left(-\frac{2n}{R^2}R^2\frac{\log(1/\alpha)}{2n}\right) \tag{G11}$$

$$= \exp[-\log(1/\alpha)] \tag{G12}$$

$$= \alpha. \tag{G13}$$

Thus the complementary event has probability at least $1-\alpha$ and implies (G8).

For the tail statement, the Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\Pr\left(\sup_z |\widehat{F}_n(z) - F(z)| > \varepsilon_n\right) \le 2e^{-2n\varepsilon_n^2} \tag{G14}$$

$$= 2\exp[-\log(2/\alpha)] \tag{G15}$$

$$= \alpha. \tag{G16}$$

On the complementary event, $F(z) \ge \widehat{F}_n(z) - \varepsilon_n$ for every z. Put $z = \widehat{q}_{1-\tau+\varepsilon_n}$. By (G9),

$$F(z) \ge \widehat{F}_n(z) - \varepsilon_n \tag{G17}$$

$$\ge 1 - \tau + \varepsilon_n - \varepsilon_n \tag{G18}$$

$$= 1 - \tau. \tag{G19}$$

Therefore $\Pr(Z > z) = 1 - F(z) \le \tau$. This workflow-cluster formulation allows arbitrary dependence among nodes within one audited workflow. □

G.6 Vector competence and scalable approximations

With R specialist states, an exact grid stores M^R competence tuples and evaluates reviewer–candidate assignments at each tuple, giving $O(TKM^R \max_t |\mathcal{P}_t| R)$ time and $O(KM^R)$ memory. When cross-skill interactions are weak, a factorized approximation

$$V_t(\mathbf{q}, b) \approx b_t(b) + \sum_{r=1}^{R} V_t^{(r)}(q^{(r)}, b) \tag{38}$$

reduces the grid cost to $O(TKRM \max_t |\mathcal{P}_t|)$; a low-rank pairwise correction can be added for selected skill pairs. These approximations preserve the scalar $R = 1$ recursion used in the paper and provide a practical scaling path beyond the two-family specialization stress test.

G.7 BELIEF-STATE COMPETENCE

When competence is not directly observed, let $\mu_t(q)$ be its posterior on the same M-point grid and let o_{t+1} be an operational observation (for example, adjudicated review accuracy or resolution time). Given transition $P(q' \mid q, x_t)$ and observation likelihood $O(o_{t+1} \mid q', x_t)$, the action-conditioned predictive belief and Bayesian update are

$$\mu^-_{t+1}(q') = \sum_q P(q' \mid q, x_t)\mu_t(q), \tag{B1}$$

$$\mu_{t+1}(q') = \frac{O(o_{t+1} \mid q', x_t)\mu^-_{t+1}(q')}{\sum_{\bar{q}} O(o_{t+1} \mid \bar{q}, x_t)\mu^-_{t+1}(\bar{q})}. \tag{B2}$$

For action $a = 0$ (skip), define $r_t(\mu, 0) = 0$; for reviewing i, define

$$r_t(\mu, i) = \sum_q \mu(q)[q\widehat{g}_{t,i} - c_{t,i}]. \tag{B3}$$

Writing $\tau(\mu, a, o)$ for (B1)–(B2), the belief-state recursion is

$$V^{\mathrm{bel}}_t(\mu, b) = \max_{a \in \mathcal{A}_t(b)} \left\{ r_t(\mu, a) + \gamma \sum_o p(o \mid \mu, a) V^{\mathrm{bel}}_{t+1}(\tau(\mu, a, o), b - \mathbf{1}[a \neq 0]) \right\}. \tag{B4}$$

This is the standard sufficient-state construction for partial observability (Chen et al., 2022); it changes the scheduler state, not HIVE's protection estimator.

Proposition 13 (Point-belief reduction)**.** *If $\mu_t = \delta_{q_t}$, the transition is deterministic $P(q' \mid q, x) = \mathbf{1}[q' = F(q, x, d_t)]$, and the next state is observed exactly, then Eq. (B4) equals Eq. 13 at every period.*

Proof. Substituting $\mu_t = \delta_{q_t}$ into (B3) removes every term except $q = q_t$:

$$r_t(\delta_{q_t}, i) = q_t\widehat{g}_{t,i} - c_{t,i} = \widehat{R}_{t,i}(q_t). \tag{B5}$$

The deterministic kernel in (B1) gives $\mu^-_{t+1} = \delta_{F(q_t, x_t, d_t)}$. Exact observation leaves that point mass unchanged in (B2). Hence the observation sum in (B4) has one nonzero next belief, and its continuation value is

$$\gamma V^{\mathrm{bel}}_{t+1}(\delta_{F(q_t, x_t, d_t)}, b - \mathbf{1}[x_t = 1]). \tag{B6}$$

At $T + 1$, both recursions equal zero. Assuming equality at $t + 1$, (B5)–(B6) make every skip and review branch in (B4) identical to the corresponding branch of Eq. 13; taking the maximum preserves equality. Backward induction proves the claim. □

PART III: ROBUSTNESS AND DEPLOYMENT DIAGNOSTICS

H ADDITIONAL ROBUSTNESS ANALYSES

This appendix contains analyses that use the same benchmark workflows and frozen checkpoints as the main experiments. They make the operating-point and repeated-review behavior explicit while leaving the primary one-review estimand unchanged.

H.1 RETENTION-WEIGHT SENSITIVITY

We vary the coefficient multiplying the competence-retention term while keeping the HIVE-L seed-11 predictions fixed. Targets are recomputed from the published benchmark equation, and pooled capacity-8 allocation is repeated 100 times. This isolates how the learned ordering behaves as the relative importance of capability retention changes.

The ranking remains useful across the tested range; increasing the retention weight increases attainable utility and reduces Source-A regret, while Source-B regret is stable. The analysis is intentionally a frozen-policy stress test that complements retraining results at each operating point.

Table 14: Frozen-policy sensitivity to retention coefficient.

Coefficient	A utility	A regret	B utility	B regret
0.025	.634	.050	.612	.038
0.050	.722	.035	.727	.036
0.075	.809	.030	.842	.034

I ONLINE-ARRIVAL AND GRAPH-QUALITY STRESS TESTS

The full-queue dynamic program in the main text is an upper-information reference that observes the realized episode queue. We additionally implemented a receding-horizon controller driven by the current pool and replanned after every period. Future pools are forecast from the remaining test workflows, with probability ν of dropping each forecasted pool. Its information set comprises observable pools and forecasts. Table 15 reports regret over 100 paired episodes (capacity four).

Table 15: Receding-horizon HIVE-DP under noisy future-arrival forecasts.

Source	$\nu = 0$	$\nu = .25$	$\nu = .50$
TheAgentCompany-style	$.0219 \pm .0174$	$.0254 \pm .0203$	$.0390 \pm .0293$
WorkArena-style	$.0136 \pm .0094$	$.0195 \pm .0156$	$.0350 \pm .0242$

To quantify sensitivity to structural graph errors, we randomly removed a fraction of observed edges at inference time while keeping the checkpoint and true paired outcomes fixed. At capacity eight, HIVE regret changes from .0510 with the intact graph to .0743, .0777 and .0579 when 10%, 25% and 40% of edges are removed (MAE .0769, .0779, .0834 and .0852 respectively). The non-monotonic tail reflects sparse graphs in which removing a redundant edge can occasionally reduce an over-propagation error. At 10/25/40% corruption, edge-type relabelling gives regret .0623/.0706/.0699 and adding spurious forward edges gives .0516/.0630/.0640. Across all three mechanisms, structural corruption degrades calibration or allocation, with missing true descendants causing the largest peak regret.

J CALIBRATION AND SELECTIVE-RISK ANALYSIS

On the held-out long-horizon test set, ten equal-frequency bins have absolute prediction–target gaps ranging from .004 in the central bin to .185 in the lowest-valued bin. Selecting the top 10%, 20%, 30%, 50%, 80% and 100% of candidates by predicted value gives MAE .1326, .1028, .0763, .0557, .0602 and .0788; the corresponding mean realized values are .0896, .0756, .0673, .0435, .0237 and .0139. The model is therefore most useful as a selective ranking signal under a tight budget; raw scores should be recalibrated before being interpreted as probabilities or absolute monetary values.

J.1 CAPACITY-CUTOFF AUDIT

Because allocation depends most strongly on candidates near the Kth score, we audit that boundary directly alongside global MAE. From the frozen seed-11 test predictions, we repeatedly draw 16 of the 24 held-out workflows per source and allocate $K = 8$. A boundary candidate is the highest-scored candidate in its workflow whose prediction lies within one half of the pool score standard deviation of the predicted eighth score. Table 16 reports 500 overlapping pool draws; pool standard deviations describe draw-to-draw variation, while candidate occurrences are not treated as independent observations.

Table 16: Selection-boundary audit on frozen predictions. Membership accuracy compares predicted and oracle top-K membership among boundary occurrences.

Source	Set Jaccard	Precision@K	Boundary MAE	Boundary accuracy
Source A	$.628 \pm .151$	.761	.0986	.709
Source B	$.669 \pm .121$	.795	.1104	.671

The local MAE motivates exposing the score margin at the cutoff and routing near-ties to a conservative or randomized logging policy. Set-level precision supports useful allocation while score recalibration handles applications that require absolute-value interpretation.

J.2 Auditable applicability gate

A new workflow source becomes eligible for discretionary HIVE allocation when three gates pass on a frozen calibration window. First, the intervention contract must match: REVIEW occurs before execution, retains the same restored state and frozen agent, and exposes only the allowed revised action. Second, the identification audit must pass: paired replay completion and branch-failure rates are reported; for logged-branch estimation, propensity is bounded by a declared ϵ and effective sample size is reported. Third, observable support must pass: node/edge-type counts, depth, candidate position, numeric features, and encoder distance are compared with training-cluster quantiles; the source enters the validation envelope when its pre-registered nonconformity score is at most the 95th percentile of held-out training clusters. Passing establishes comparability to that envelope. Other sources follow the neutral mandatory-review policy while additional data are collected. The graph corruptions, response shift, MiniLM swap, and pooled fallback audit instantiate individual parts of this checklist, whereas an independent workflow study must exercise the complete gate.

K Learning the competence transition from logs

The transition parameters used by HIVE-DP are published environment dynamics. In deployment they can be estimated from longitudinal oversight logs as follows. First, define competence before examining allocation outcomes, using a rolling task-relevant operational score such as adjudicated accuracy, and freeze its window and normalization on the training period. For reviewer trajectory j and time t, fit

$$q_{j,t+1} = \min\{1, q_{\min} + (q_{j,t} - q_{\min})e^{-\lambda d_{j,t}} + \eta x_{j,t}(1 - q_{j,t})\} + \varepsilon_{j,t+1} \tag{T1}$$

by constrained nonlinear least squares with $q_{\min} \in [0, 1]$, $\lambda \geq 0$, and $\eta \in [0, 1]$. Splits are by reviewer, or strictly forward in time when the same reviewer must occur in both periods, keeping adjacent samples from one trajectory within a single fit or calibration partition.

Second, resample whole reviewer trajectories and refit (T1) in each of B cluster-bootstrap replicates. Let I_q, I_λ, I_η be the resulting percentile intervals. Their Cartesian product is the rectangular transition set used by the robust planner. Since future value is nondecreasing in competence under Section 3's conditions, its conservative endpoint uses the largest admissible decay and smallest admissible retention; $q_{\min}$ is chosen by direct minimization over I_q. The intervals and controller are frozen before test episodes.

Third, admit the fit when both $x = 0$ and $x = 1$ occur in every declared task family, idle-duration support covers the deployment horizon, and held-out one-step residual coverage meets its pre-registered level. These support checks establish the predictive use of (T1); causal use additionally requires review assignment to be independent of latent reviewer state conditional on the model inputs. The neutral mandatory-review policy covers operating points awaiting additional transition evidence. The current benchmark reports correctly specified dynamics in the main table, with the parameter-shift audit serving as a pre-specified robustness test.

Finite-log transition audit. We additionally test the complete fitting-to-planning path. We sample 64 nested reviewer trajectories of length 12 from the published nominal transition, randomize review with probability .5, draw idle exposure uniformly from $[.02, .10]$, and add Gaussian observation noise with standard deviation .01. Fits use the first $n \in \{8, 24, 64\}$ trajectories; 200 whole-trajectory bootstrap samples form the robust set. One-step RMSE is measured on 32 unseen reviewer trajectories. The frozen seed-11 HIVE scores are then scheduled over the same 100 held-out workflow episodes per source at capacity four. Table 17 reports regret against the true-outcome, true-transition oracle.

At 64 trajectories, the fitted floor/decay/retention are $.248/1.802/.081$ versus $.25/1.8/.08$, and learned-transition regret matches the known-transition schedule to four decimals on both sources. Robust planning improves Source-A regret, while Source B favors different operating points across n; Section 3 therefore selects the controller by validation performance. With reproducible reviewer

Table 17: Finite-log transition audit. Known-transition HIVE-DP regret is .0228/.0144 on Sources A/B. Lower is better.

Trajectories	Held-out RMSE	A learned	A robust	B learned	B robust
8	.0101	.0227	.0191	.0147	.0149
24	.0100	.0228	.0206	.0147	.0144
64	.0099	.0228	.0216	.0144	.0142

logs and a declared transition law, this audit establishes the finite-sample behavior of the full pipeline and provides the protocol for subsequent human skill-retention validation.

L OBSERVABLE REVIEW-OVERHEAD SENSITIVITY

We vary the per-review overhead supplied to both the planner and the oracle from 0 to $.05$ while retaining the frozen HIVE predictions and capacity four. Over 100 episode orderings, company-operations regret is $.0237/.0238/.0216/.0202$ and service-operations regret is $.0136/.0138/.0130/.0125$ for overhead $0/.01/.03/.05$, respectively (standard deviations are reported in the run ledger). The stable ordering across positive costs shows that the scheduler responds to overhead through the skip/review threshold while keeping cost separate from protection gain.

M REVIEWER SPECIALIZATION STRESS TEST

The main benchmark uses a scalar competence state because every review option is mapped to a common standardized contract. To examine the complementary specialist setting, we instantiate a two-family stress test in which competence is family-specific and a review only renews the corresponding family. A vector dynamic program observes $(q^{(1)}, q^{(2)})$, whereas a scalar policy ranks the current candidate using their mean. Over 500 sixteen-period episodes with capacity four, the vector policy's oracle utility is .4573 and the scalar policy's utility is .3568 (regret .1005 $\pm$ .0949). This result motivates a vector competence state and reviewer–task assignment matrix for specialist settings, while the scalar model serves standardized review contracts.

M.1 FACTORIZED COMPETENCE EVALUATION

To test the practical scaling direction in the vector-state derivation, we added a calibrated factorized controller to the two-family stress test. The controller keeps the family-specific competence coordinates but scores the current option with $q^{(r)} g_i$; it ignores cross-family interactions in its value approximation. A scalar comparator uses $\bar{q} g_i$. The review threshold is selected on 250 calibration episodes from a grid of 29 values in $[.04, .18]$, and the frozen thresholds are evaluated on 250 new episodes. The vector oracle still uses the exact two-dimensional Bellman recursion.

Table 18: Calibrated factorized competence controller.

Controller	Utility	Regret	Regret sd
Vector oracle	.4559	–	–
Scalar competence	.3937	.0622	.0614
Factorized family-aware	.4229	.0330	.0410

The family-aware approximation closes 46.9% of the scalar controller's regret gap while retaining a simpler state representation than the exact joint grid. This evaluation positions factorization as a practical middle ground when the number of specialist states makes M^R enumeration unattractive.

M.2 RECEDING-HORIZON EXECUTION

At each period t, the online controller receives the current candidate pool $\mathcal{P}_t$ and a forecast $\widehat{\mathcal{P}}_{t+1:T}$. It constructs the corresponding finite-horizon DP, executes only its first action, observes the next pool, updates q_t using the realized review indicator, and repeats. In pseudocode:

1. Observe $(q_t, b_t, \mathcal{P}_t)$ and forecast future pools.

2. Solve Eq. 13 on the forecast horizon.
3. Execute $a_t = \arg\max\{\textsf{skip}, \textsf{review}(i) : i \in \mathcal{P}_t^+(q_t)\}$.
4. Set $q_{t+1} = F(q_t, a_t, d_t)$ and $b_{t+1} = b_t - \mathbf{1}[a_t \neq \textsf{skip}]$.

The controller uses the observable arrival forecast at each replanning step; future outcomes and benchmark-construction metadata remain evaluation-side quantities.

N UNCERTAINTY-AWARE SCHEDULING

We form an ensemble from the three frozen HIVE checkpoints (seeds 11, 17, and 23) and use $\widetilde{g} = \bar{g} - \kappa\sigma$ in a robust HIVE-DP. For $\kappa > 0$ the planner also uses the worst admissible transition in the set $\rho \in [1.8, 2.2]$ and $\eta \in [.04, .08]$. Over 100 nominal episodes at capacity four, regret for $\kappa \in \{0, .5, 1\}$ is:

Table 19: Lower-confidence HIVE-DP sensitivity.

Source	$\kappa = 0$	$\kappa = .5$	$\kappa = 1$
TheAgentCompany-style	.0148 ± .0158	.0177 ± .0176	.0296 ± .0235
WorkArena-style	.0172 ± .0127	.0150 ± .0109	**.0139 ± .0099**

The calibrated risk parameter is therefore source-dependent: conservative disagreement penalties benefit the noisier WorkArena-style source, while the company-operations source favors the neutral setting. This source dependence makes validation an informative deployment diagnostic. If the propagated total-value error of every candidate is bounded by $\epsilon_{t,i} = \kappa\sigma_{t,i}$, then for any two schedules with at most K reviews the standard perturbation argument gives $|U(\pi) - \widetilde{U}(\pi)| \leq \sum_{(t,i)\in\pi} \epsilon_{t,i}$ and hence $U(\pi^*) - U(\widetilde{\pi}) \leq 2\max_{\pi:|\pi|\leq K} \sum_{(t,i)\in\pi} \epsilon_{t,i}$. The bound treats transition dynamics as fixed and makes explicit how ensemble disagreement controls worst-case allocation regret.

O REPEATED-REVIEW STRESS TEST

The main estimand permits at most one reviewed node per workflow. To quantify the behavior when a queue allows two reviews, we use the same seed-11 test predictions and set the per-workflow cap to two while retaining total capacity $K = 8$. With unit review costs and additive paired outcomes, the exact allocator selects the highest-scoring feasible candidates.

Table 20: Two-review-per-workflow stress test (seed 11, capacity 8).

Configuration	A utility	A regret	B utility	B regret
HIVE-L	.722	.035	.784	.039
no horizon state	.722	.035	.784	.039

The stress test shows that both estimators remain operational when repeated reviews are admitted; their seed-11 schedules coincide at this capacity. The dynamic experiment below evaluates temporal coupling directly, while this stress test isolates the feasibility constraint for multiple nodes in one workflow.

Proposition 14 (Exact repeated-review allocation under modular values)**.** *Let each candidate i have a fixed value w_i, let total capacity be K, and let workflow r admit at most k_r selected candidates. The algorithm that scans candidates in nonincreasing w_i, accepting a candidate exactly when total and workflow capacity remain, returns an optimal set for*

$$\max_{S} \sum_{i\in S} w_i \quad \textit{s.t.} \quad |S| \leq K, \quad |S \cap \mathcal{I}_r| \leq k_r \textit{ for every } r. \tag{R1}$$

Thus the stress-test allocator is exact when paired outcomes are additive and history-independent. If reviewing one node changes another node's marginal value, exactness is instead recovered by augmenting the state with review history and solving the corresponding Bellman recursion.

Proof. Discard candidates with $w_i \leq 0$, since removing any such candidate preserves feasibility and cannot decrease the objective. Let the greedy set be $G = \{g_1, \ldots, g_m\}$ in acceptance order, where $m \leq K$. Take an optimal set O of maximum cardinality among objective-maximizing sets. We construct optimal sets O_j containing $\{g_1, \ldots, g_j\}$.

Set $O_0 = O$. Suppose O_{j-1} has been constructed. If $g_j \in O_{j-1}$, set $O_j = O_{j-1}$. Otherwise consider the moment greedy accepted g_j. All $g_1, \ldots, g_{j-1}$ are already in O_{j-1}. If $|O_{j-1}| < K$ and workflow $r(g_j)$ is below its cap in O_{j-1}, then $O_{j-1} \cup \{g_j\}$ is feasible and has at least as much value and larger cardinality, contradicting the choice of O; hence one constraint must bind.

If workflow $r(g_j)$ binds, that workflow contains an element $o \in O_{j-1} \setminus \{g_1, \ldots, g_{j-1}\}$. If only total capacity binds, there is likewise some $o \in O_{j-1} \setminus \{g_1, \ldots, g_{j-1}\}$. In either case define

$$O_j = (O_{j-1} \setminus \{o\}) \cup \{g_j\}. \tag{R2}$$

The replacement preserves the total cap; in the workflow-binding case it also preserves that workflow count, and in the total-only case the destination workflow had spare capacity. Hence O_j is feasible. Because greedy scans in nonincreasing order and o was not among the earlier accepted elements, $w_{g_j} \geq w_o$. Therefore

$$\sum_{i \in O_j} w_i = \sum_{i \in O_{j-1}} w_i - w_o + w_{g_j} \tag{R3}$$

$$\geq \sum_{i \in O_{j-1}} w_i. \tag{R4}$$

Since O_{j-1} is optimal, (R4) implies that O_j is also optimal. Induction produces an optimal set containing all of G. If it contained an additional positive-valued feasible candidate, greedy would have accepted that candidate before terminating; hence no such addition exists and G itself is optimal.

When marginal values depend on prior reviews, the equality $U(S) = \sum_{i \in S} w_i$ used in (R3) is unavailable. Adding the history to the state restores the Markov property, after which exhaustive Bellman maximization is exact by backward induction. □

P Dynamic-transition robustness

We repeat the 16-period scheduler experiment under faster competence decay ($d_t = .08$) and weaker review retention ($\eta = .04$), changing one transition component at a time. Table 21 reports capacity-4 regret over seven estimator seeds and 100 episode orderings per seed.

Table 21: Dynamic scheduler robustness across competence transitions.

Regime	Policy	Source A	Source B
Rapid decay	no-horizon static	.0542	.0420
	HIVE-L one-step	.0541	.0424
	HIVE-DP	**.0222**	**.0191**
Weak retention	no-horizon static	.0490	.0380
	HIVE-L one-step	.0489	.0384
	HIVE-DP	**.0241**	**.0203**

Relative to one-step HIVE-L, HIVE-DP lowers regret by 59.0%/55.0% under rapid decay and 50.7%/47.1% under weak retention for Sources A/B. Under the nominal transition the corresponding reductions are 51.3%/48.4% at capacity four and 46.3%/32.0% at capacity six. The planner therefore benefits most when the timing of practice has a stronger effect on future capability, while retaining its advantage when retention is reduced.

P.1 Misspecified retention shape

We additionally alter the environment in a way that is absent from the planner: HIVE-DP plans with the linear reviewed-state increment $\eta(1 - q_t)$, while both the realized transition and the oracle use $\eta(1 - q_t)^2$. Thus the learned scores, planning model, and policy are frozen, with the test-time

Table 22: Regret under an unmodeled quadratic retention transition, mean (SD).

Policy	Source A	Source B
no-horizon static	.0478 (.0382)	.0372 (.0253)
HIVE-L one-step	.0477 (.0383)	.0376 (.0259)
HIVE-DP	**.0238** (.0250)	**.0199** (.0146)

retention law changed. Table 22 reports the same seven seeds and 100 shared orderings per seed at capacity four.

Despite planning with the linear law, HIVE-DP lowers regret relative to HIVE-L by 50.1% and 47.1% on Sources A and B. This intervention isolates transition shape misspecification from estimator retraining and complements the parameter-shift tests above.

P.2 Misspecified competence–reward link

Equation 13 assumes that current protection scales linearly as $q_t g_{t,i}$. To make this assumption falsifiable, we freeze the linear HIVE-DP policy and evaluate its actions when realized reward instead follows $q_t^{\alpha} g_{t,i} - c_{t,i}$. The oracle knows α, whereas HIVE-DP continues to plan with $\alpha = 1$; neither estimator scores nor transitions are refitted. Table 23 uses capacity four, three estimator seeds, and 100 episode orderings per seed and source. The same finite workflows recur across orderings, so the standard deviations describe episode variation rather than additional independent task evidence.

Table 23: Regret under a misspecified competence–reward exponent, mean (standard deviation). HIVE-DP always plans with exponent one.

		Source A		Source B	
α	Policy	Regret	Reduction	Regret	Reduction
.5	HIVE-L one-step	.0459 (.0332)	–	.0373 (.0314)	–
	linear HIVE-DP	.0313 (.0309)	31.8%	.0244 (.0196)	34.7%
1.0	HIVE-L one-step	.0432 (.0274)	–	.0355 (.0267)	–
	linear HIVE-DP	.0221 (.0205)	49.0%	.0163 (.0132)	54.1%
1.5	HIVE-L one-step	.0399 (.0232)	–	.0337 (.0234)	–
	linear HIVE-DP	.0175 (.0154)	56.2%	.0132 (.0104)	60.8%
2.0	HIVE-L one-step	.0361 (.0204)	–	.0309 (.0209)	–
	linear HIVE-DP	.0152 (.0131)	57.9%	.0117 (.0090)	62.2%

The frozen linear planner retains lower regret than one-step allocation across this evaluated range, with a smaller advantage for the concave $\alpha = .5$ link. The stress test evaluates planner sensitivity across candidate links; longitudinal validation selects the link and retains the dynamic controller when it improves held-out likelihood or decision regret over the static fallback.

Q Transition-set stress grid

The transition-set planner was evaluated on a prespecified 3×3 grid of actual decay and retention values spanning the nominal point. The risk-neutral policy uses $(\rho, \eta) = (1.8, .08)$ in planning; the robust policy uses $(2.2, .04)$ together with $\kappa = 1$. Both schedules are then replayed under each actual transition and compared with an oracle that knows that transition. Values are mean regret over 100 paired episode orderings at capacity four; "gain" is the percentage reduction of robust regret relative to risk-neutral (positive is better).

The grid makes the operating condition explicit: transition robustness is beneficial for Source B across the tested region, whereas the conservative gain penalty is too costly for Source A. We therefore treat κ and the transition radius as validation-calibrated deployment knobs: the selector retains the nominal risk-neutral setting unless validation supports a robust choice.

The selector uses 30 validation orderings and the one-standard-error rule; the grid averages aggregate the nine transition settings and 100 held-out orderings per setting. Relative to the corresponding neutral grid averages (.0146 and .0168), the calibrated controller preserves Source-A performance and reduces Source-B regret by 22.9%. This audit is also a guard against choosing a conservative transition set solely because it is pessimistic.

Table 24: Actual-transition grid for Source A (company-operations motifs).

ρ	η	Neutral	Robust	Gain (%)
1.8	.04	.0143	.0277	-94.4
1.8	.06	.0145	.0287	-97.5
1.8	.08	.0148	.0296	-100.3
2.0	.04	.0143	.0277	-94.1
2.0	.06	.0145	.0287	-97.5
2.0	.08	.0148	.0296	-100.5
2.2	.04	.0144	.0278	-93.3
2.2	.06	.0146	.0288	-97.0
2.2	.08	.0148	.0297	-100.2

Table 25: Actual-transition grid for Source B (service-operations motifs).

ρ	η	Neutral	Robust	Gain (%)
1.8	.04	.0169	.0134	20.5
1.8	.06	.0170	.0136	20.0
1.8	.08	.0172	.0139	19.3
2.0	.04	.0166	.0134	19.7
2.0	.06	.0168	.0136	19.1
2.0	.08	.0171	.0139	18.4
2.2	.04	.0164	.0133	19.0
2.2	.06	.0166	.0136	18.3
2.2	.08	.0169	.0139	17.7

Table 26: Validation-calibrated operating points and held-out grid averages.

Source	Selected controller	Validation regret	Grid regret
Source A	neutral	.0111	$.0157 \pm .0145$
Source B	LCB nominal ($\kappa = .5$)	.0212	$.0130 \pm .0104$

Q.1 Pooled fallback audit

For sources represented in the pooled validation set, the implementation applies the same one-standard-error selector to that pool. We ran this fallback with 30 pooled validation orderings and then froze the selected controller before evaluating 100 test orderings per source. The selector retained the neutral controller: validation regret was $.01529 \pm .00240$ for neutral, while the best nominal lower-confidence mean was $.01348$ and remained within one standard error. On the test episodes, the frozen pooled choice achieved regret $.01369 \pm .01304$ for Source A and $.01444 \pm .01072$ for Source B.

Initial-competence heterogeneity. Under nominal capacity-4 dynamics, HIVE-DP versus one-step regret at initial competence $.45/.60/.75$ is respectively $.0221/.0253/.0319$ versus $.0207/.0280/.0425$ on Source A, and $.0205/.0215/.0290$ versus $.0228/.0267/.0422$ on Source B. The dynamic planner improves five of the six source–state combinations and yields its largest gains when more competence is available to preserve over the episode.

Part IV: Detailed workflow cases

R Detailed workflow case studies

These worked cases expose the decision semantics hidden by aggregate metrics. They reconstruct held-out workflow motifs with the paper's scoring equations; values are rounded from recorded benchmark fields. These reproducible benchmark cases instantiate the modeled decision semantics and evaluation quantities.

Table 27: Pooled validation fallback (capacity four).

Source	Selected policy	Test regret (mean ± sd)
Source A	neutral	.01369 ± .01304
Source B	neutral	.01444 ± .01072

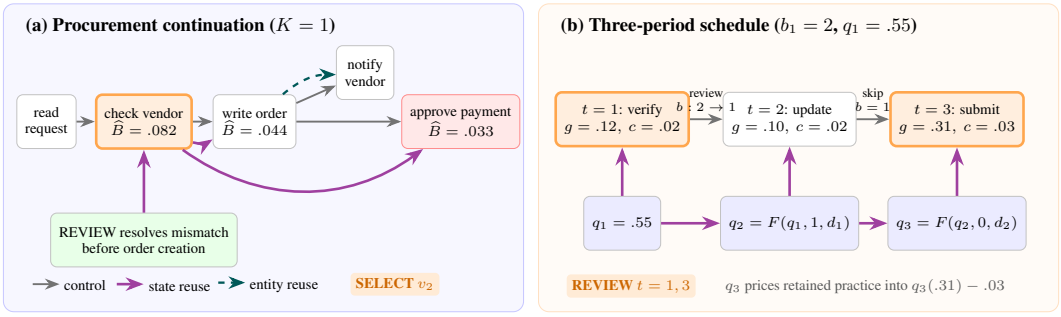


Figure 2: Worked cases for HIVE and HIVE-DP. (a) The violet state paths show why correcting vendor identity at v_2 changes both order creation and payment; the dashed entity path carries the written order to notification. (b) The selected review at $t = 1$ changes q_2; after deferring at $t = 2$, HIVE-DP values the terminal submission using the resulting q_3. Orange borders mark selected reviews.

R.1 Case 1: upstream verification dominates a high-risk terminal action

A procurement workflow asks an agent to validate a supplier, create a purchase order, notify the supplier, and submit payment for approval. Its five nodes are v_1 =ReadRequest, v_2 =CheckVendor, v_3 =WriteOrder, v_4 =NotifyVendor, and v_5 =ApprovePayment. Control edges are $v_1 \to v_2 \to v_3 \to v_5$ and $v_3 \to v_4$. State edges connect v_2 to v_3 and v_5 because the verified vendor identifier is reused; an entity edge connects v_3 to v_4 through the purchase-order object. The factual trace contains an ambiguous vendor match at v_2. AUTO retains the top retrieved vendor, whereas the hypothetical-review mode asks the reviewer to accept, revise, or block that match before any order is written.

At allocation time HIVE sees the prefix through the proposed v_2 action, the typed graph, both frozen mode representations, and numeric features. The later revision and paired outcome remain beyond this information boundary. The batch also contains candidates at v_3 and v_5, but capacity permits one review. Table 28 decomposes their recorded estimates. Net value is local delta plus propagated delta minus review overhead; the respective overheads are .026, .021, and .019.

Table 28: Procurement case. Values are net protection estimates at observed competence $q = .60$; higher is better.

Candidate	local delta	propagated delta	net value
CheckVendor	.021	.087	.082
WriteOrder	.034	.031	.044
ApprovePayment	.052	.000	.033

The arithmetic gives $.021 + .087 - .026 = .082$ for CheckVendor, $.034 + .031 - .021 = .044$ for WriteOrder, and $.052 + 0 - .019 = .033$ for ApprovePayment. A confidence or terminal-risk rule selects v_5 because it is externally consequential and has the largest local delta. HIVE selects v_2: resolving vendor identity changes the order, notification, and payment record before they become externally visible. Under REVIEW, a mismatch is blocked and the continuation is regenerated from the corrected state; under AUTO, the same mistake is copied into three descendants. The candidate-local ablation removes the .087 propagated component and reverses the ranking, while the untyped-graph ablation weakens the state link from verification to payment. The decision therefore follows the concrete continuation effect represented by the typed dependencies.

The case also clarifies the zero-intervention invariant. If AUTO and REVIEW have identical representations at v_2, their centered delta is zero, every propagated message remains zero, and HIVE returns

zero intervention value even for a large factual graph. Positive review scores therefore originate from the intervention delta.

R.2 Case 2: deferring review preserves budget but loses competence

Consider a three-period queue with total budget two and initial competence $q_1 = .55$. Period 1 contains vendor verification with normalized protection gain .12 and cost .02; period 2 contains an internal record update with gain .10 and cost .02; period 3 contains a rare external submission with gain .31 and cost .03. The arrival order, idle exposures, and costs are known to the planner, while paired outcomes remain hidden. At a common competence .55, the immediate net rewards are .046, .035, and .1405. The last action is clearly valuable; the dynamic recursion additionally determines which earlier action should supply a practice opportunity.

HIVE-DP initializes $\widehat{V}_4(q, b) = 0$ for every grid state and works backward. At period 3 it compares zero with $q(.31) - .03$ whenever $b > 0$. At period 2 it compares the skipped state $F(q, 0, d_2)$ with the reviewed state $F(q, 1, d_2)$ and decrements budget only in the latter branch. At period 1 the two complete branch values are

$$Q_{\text{skip}} = .97\widehat{V}_2(F(q_1, 0, d_1), 2), \tag{39}$$

$$Q_{\text{review}} = .55(.12) - .02 + .97\widehat{V}_2(F(q_1, 1, d_1), 1). \tag{40}$$

The .046 first term in Q_{review} is immediate protection. Every later gain is then evaluated at its branch-specific competence. Reviewing period 1 raises the competence passed to periods 2–3 but consumes one of two units; skipping preserves both units but permits decay. Backward evaluation reviews periods 1 and 3 for this recorded queue and skips the lower-value internal update. A one-step rule selects the same pair here for a different reason; perturbing the period-1 gain around its decision boundary separates the policies because only HIVE-DP includes future capability.

Two targeted changes identify this term. Setting retention $\eta = 0$ makes $F(q, 1, d) = F(q, 0, d)$, so the competence advantage vanishes exactly and only immediate reward plus budget remains. Removing the period-3 pool sharply reduces the value of retained competence because no high-gain future action can use it. Shuffling the same three arrivals changes which review provides practice before the submission while holding the multiset of gains fixed. These changes are the mechanisms tested at scale by the no-retention and shuffled-arrival controls: the dynamic advantage tracks temporal opportunity while candidate scores and forecast features remain fixed.