WHEN CONFIDENCE IS NOT ENOUGH: ACTION-CONTRASTIVE SELF-MODELS UNDER CAPABILITY DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

A calibrated agent may estimate its probability of success accurately and still choose the wrong recovery. Retrieval, verification, replanning, escalation, and abstention repair different failures, so scalar confidence removes the direction needed for action choice. We formulate self-knowledge as an action-contrast surface: the change in terminal value induced by each available recovery branch. SCOPE factors this surface through learned operational response modes, aligns their action–diagnostic signatures after a model, tool, access, or context change, and routes by predicted net value. Under conditional response homogeneity, $G = Db$ is sufficient for one-step recovery choice. In a multi-regime linearization, sharing K modes across R regimes changes the stylized parameter count from RAd to $K(d + RA)$, while signature separation gives a condition for transporting persistent modes across versions. We evaluate these claims with real local-model and tool executions: every legal branch is run from the same reset pre-action state, terminal outcomes are scored by task-native evaluators, and only one logged branch outcome is exposed to the learner. Across 6 open-weight checkpoints and six transition families, action conditioning raises mean normalized utility from .681 to .718, generic rank-matched sharing reaches .725, and SCOPE reaches .731. SCOPE further reduces held-out intervention-gain error by 21–26% relative to the rank-matched control. Direct-AV catches up with abundant feedback and achieves higher utility when new response mechanisms or noisy diagnostics break the shared geometry.

1 INTRODUCTION

Two tasks can receive the same predicted success probability and require opposite actions. Suppose an agent assigns both tasks confidence 0.62. In the first, the needed fact is absent but retrievable; in the second, the fact is present but the agent lacks permission to act. Retrieval is valuable in the first case and nearly useless in the second, while escalation has the reverse pattern. A scalar can support answer-versus-abstain, but it does not determine which of several costly recovery policies will repair the current failure. Figure 1 illustrates this distinction; the displayed confidences and mode weights are schematic.

Capability change makes this distinction operationally important. Public model services and updated checkpoints can change behavior across release snapshots, and instance-level regressions can coexist with aggregate improvements (Chen et al., 2024; Echterhoff et al., 2024). A model update may reduce the frequency of reasoning failures while simultaneously making an attached verifier less reliable. Aggregate accuracy can remain nearly unchanged although the best recovery switches from verification to replanning. We call the first movement incidence drift—how often operational response modes occur—and the second efficacy drift—what an action does within a mode. A useful self-model must distinguish these movements before paying for retrieval, extra computation, tools, or human escalation.

Why not directly predict the value of every action? This is the strongest alternative and, with sufficient post-change feedback, often the right one. The difficulty is that deployment logs expose only the branch that was selected, some actions receive little support, and related model–tool–task

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.

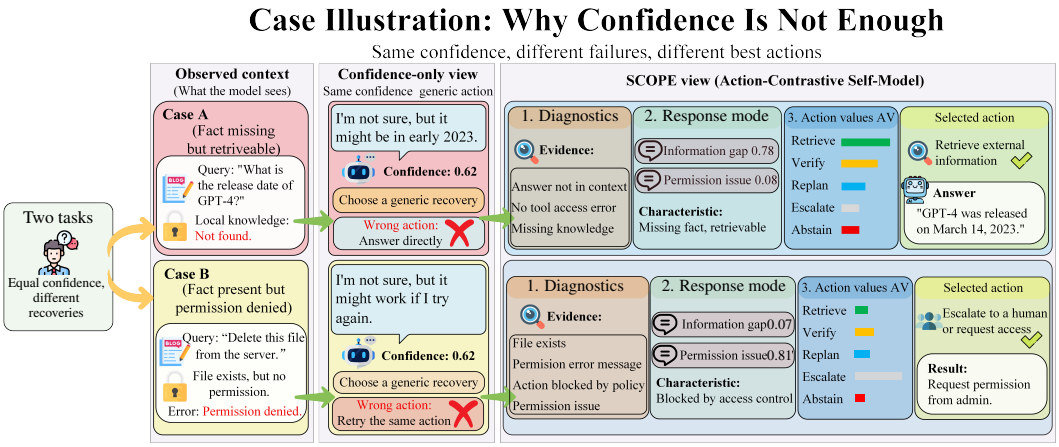


Figure 1: Equal confidence can require different recovery. A retrievable knowledge gap favors retrieval, whereas a permission gap favors escalation. The mode weights and value bars illustrate the decision mechanism; they are not measured experimental outcomes.

combinations may share recurring response structure. Logged-bandit and treatment-effect methods provide the correct statistical framing, but they do not by themselves decide which structure should be shared. SCOPE represents a task as soft mass over operational response modes and each mode by its vector of branch effects. The representation can pool evidence across contexts and capability regimes, but it can also fail when modes merge, split, or newly emerge. We therefore separate three comparisons: scalar confidence versus action values, unrestricted action values versus generic low-rank sharing, and generic sharing versus SCOPE's diagnostic modes and cross-version alignment.

We make three contributions. (i) We characterize the action contrasts sufficient for one-step recovery and identify when confidence loses a decision direction. (ii) We study shared response geometry across capability regimes and give a separation condition for transporting persistent modes. (iii) We separate action conditioning, generic sharing, and operational response structure, using selected-branch training feedback and paired branch evaluation. The comparisons include regimes where the structure fails: abundant feedback, new response mechanisms, weak support, and corrupted diagnostics.

2 Related Work

Confidence, abstention, and selective action. Query-level uncertainty can predict answerability before generation and trigger retrieval or cascades (Chen et al., 2026). Verbal confidence can be calibrated (Tian et al., 2023), selective generation controls answer risk (Lee et al., 2024), and reasoning traces can improve abstention (Gourabathina et al., 2026). Tool-oriented work calibrates prompt or trajectory choices (Liu et al., 2024a), distinguishes evidence from verification signals (Xuan et al., 2026), and attributes tool failures to candidate repairs (Wang et al., 2026). These methods establish that uncertainty and diagnostics can guide behavior. Our question begins where a scalar ends: several interventions may be legal, costly, and differently effective even when answerability is perfectly calibrated.

Action values, treatment effects, and logged feedback. Direct regression, inverse propensity weighting, and doubly robust estimation are standard foundations for learning policies from partially observed action outcomes (Dudík et al., 2011; Swaminathan & Joachims, 2015; Wang et al., 2017). Estimating action contrasts can be statistically simpler than modeling complete outcome surfaces when treatment effects have lower complexity (Carranza et al., 2023). Related multi-task off-policy work shares information across similar logged-bandit tasks (Hong et al., 2023). SCOPE does not replace these estimators: Direct-AV is the unrestricted regression alternative in every experiment, and the outcome loss is propensity corrected. The narrower claim is that capability regimes

sometimes reuse a compact action-response geometry that can be estimated more efficiently than independent action heads.

Interactive agent evaluation and capability change. AgentBench, WorkArena, ToolSandbox, and τ-bench evaluate language agents through real or executable interactions rather than answer text alone (Liu et al., 2024b; Drouin et al., 2024; Lu et al., 2025; Yao et al., 2025). Longitudinal evaluations further show that the behavior of a named model service can change across time (Chen et al., 2024), while compatibility methods study how to preserve behavior across model updates (Echterhoff et al., 2024). Our focus is different: we do not rank providers or preserve the old answer. We ask which fixed recovery branch has current value after the model, tool, permission, or context regime changes, and we evaluate that choice using realized terminal outcomes.

Shared and latent decision structure. Low-rank representation learning can improve sample efficiency across contextual bandit tasks (Lin et al., 2024), and latent-state bandits maintain beliefs over hidden dynamics (Nelson et al., 2022). Predictive state representations define state through action-conditional future observations rather than semantic labels (Littman et al., 2001). Stateful policy learning and latent-context transfer show both the value and danger of transporting response structure (Kallus & Zhou, 2022; Deng et al., 2025). Value-of-information methods formalize whether an observation is worth its cost (van Merwijk et al., 2022). SCOPE specializes these perspectives to one-step recovery: its state is an operational basis for branch effects, and its central boundary is whether that basis remains supported and alignable after the transition.

3 Action-Relevant Self-Knowledge

3.1 From Confidence to Rescue Contrasts

At episode t, let $I_t = (x_t, h_t, o_t)$ denote the information available before any recovery, let v_t denote the current capability regime, and let $\mathcal{A}_t$ be the legal branch set. Branches include answer, abstain, retrieve, verify, escalate, and replan. Each branch is a fixed terminal policy, so resetting the environment to the same pre-action state makes its terminal success $Y_t^a \in \{0, 1\}$ well defined. With non-success utility $u_t(a)$, branch cost $k_t(a)$, and cost weight λ,

$$V_t(a \mid I_t, v_t) = \mathbb{E}[Y_t^a \mid I_t, v_t] + u_t(a) - \lambda k_t(a). \tag{1}$$

Let $S_t \in \{1, \ldots, K\}$ be a learned soft response mode, including an `ok` component, and define

$$\begin{aligned} b_v(s \mid I) &= P(S = s \mid I, v), \\ M_v(a, s) &= \mathbb{E}[Y^a \mid S = s, v]. \end{aligned} \tag{2}$$

The central modeling condition is explicit:

$$\mathbb{E}[Y^a \mid S = s, I, v] = M_v(a, s). \tag{3}$$

Equation 3 is conditional response homogeneity: the mode must retain the part of context that predicts how the available branches respond. It need not preserve every semantic distinction. Two semantic causes may share a mode if no branch distinguishes them; one cause should split if it induces different rescue patterns in different contexts.

Relative to the answer branch a_0, define

$$D_v(a, s) = M_v(a, s) - M_v(a_0, s), \qquad G_v(I) = D_v b_v(I). \tag{4}$$

Columns of D_v are rescue signatures. Equation 4 is not an unconditional identity: it follows from the response-homogeneity model in Eq. 3.

Lemma 1 (decision sufficiency). For fixed one-step branches, current M_v, known offsets, and correct availability,

$$T_v(I) = \big(\mathbb{E}[Y^{a_0} \mid I, v], D_v b_v(I), u, k, \mathcal{A}\big) \tag{5}$$

determines every pairwise expected-utility contrast. For fixed D_v, two beliefs b, b' preserve all relative success contrasts iff $D_v(b - b') = 0$. Over a rich offset family, $D_v b$ is a sufficient linear representation of those contrasts, with minimal coordinate dimension characterized in Appendix B.

The proof subtracts the reference value. More generally, with $q = m_0^\top b$ and $m_0 = M(a_0, :)^\top$, confidence preserves all contrasts on the full simplex precisely when each row of D lies in $\mathrm{span}\{\mathbf{1}, m_0\}$. Under a rescue-only model, this reduces to cause-invariant pairwise rescue differences.

3.2 WHY SHARE A RESPONSE BASIS ACROSS REGIMES?

Let $\phi(I) \in \mathbb{R}^d$ be a bounded context representation and let $v \in \{1, \dots, R\}$ index model, tool, access, or context regimes. A regime-specific Direct-AV linearization learns $W_v \in \mathbb{R}^{A \times d}$. The shared-mode class uses one encoder $B \in \mathbb{R}^{K \times d}$ and regime-specific response matrices $M_v \in \mathbb{R}^{A \times K}$, so $W_v = M_v B$.

Proposition 1 (shared multi-regime estimation). Under bounded norms, independent logged samples, realizability, and minimum propensity $p_{\min} > 0$, the leading estimation terms for the shared and direct linear classes are

$$\widetilde{O}\left(\sqrt{\frac{K(d+RA)}{np_{\min}}}\right) \quad \text{and} \quad \widetilde{O}\left(\sqrt{\frac{RAd}{np_{\min}}}\right) \tag{6}$$

respectively, with additional terms of order $\widetilde{O}(p/(np_{\min}))$ for parameter count p. Appendix B states the risk normalization, action-support assumptions, and complete bound.

The comparison is deliberately cross-regime. Its benefit requires $K(d + RA) < RAd$, adequate support in each regime, and an accurate shared-basis assumption. If a separate encoder is trained for each regime, the comparison reduces to $K(A+d)$ versus Ad; no low-rank advantage is implied when K is not smaller than the effective action dimension. Misspecification adds approximation error, so Direct-AV should catch up with abundant data and can win when a new response mechanism lies outside the shared basis. The same rate applies to LowRank-AV. It describes a shared linear basis, not the full nonlinear encoder; the diagnostic objective and alignment require separate empirical evidence.

3.3 WHAT SURVIVES A CAPABILITY TRANSITION?

Response modes are aligned by what they predict, not by names. For actions supported on both sides of a transition, let $\mathcal{C}$ denote the common action set and write

$$h_v(s) = \big[M_v(\mathcal{C}, s), L_v(:, s)\big], \tag{7}$$

where L_v contains permitted diagnostic emissions. Persistent modes should have nearby standardized signatures; a newly emerged or unsupported mode should remain unmatched.

Proposition 2 (stable signature alignment). Assume persistent modes admit a correspondence with within-pair movement at most η, estimated signatures have error at most ϵ, and every incorrect or new-state match is separated by at least Δ. If $\Delta > \eta + 4\epsilon$, thresholded partial matching recovers the persistent correspondence. Any threshold in $(\eta + 2\epsilon, \Delta - 2\epsilon)$ accepts persistent matches and rejects unmatched modes.

Correct pairs are at estimated distance at most $\eta + 2\epsilon$; incorrect pairs are at least $\Delta - 2\epsilon$. Matching allows unmatched modes and assumes consistently estimated signatures in common coordinates. When a merge, split, support loss, or diagnostic corruption closes this gap, the proposition intentionally provides no guarantee.

From estimation error to routing error. Let ε_{est}, $\varepsilon_{\text{align}}$, and ε_{mis} bound supported value-estimation, alignment, and best-in-class approximation error in value units. If unsupported or misaligned events occur with probability ρ and legal action values span at most B, the standard argmax comparison gives expected one-step routing regret at most

$$2(\varepsilon_{\text{est}} + \varepsilon_{\text{align}} + \varepsilon_{\text{mis}}) + B\rho. \tag{8}$$

This is an accounting bound, not an identifiability theorem for arbitrary logs. Positive action support and response homogeneity control estimation; signature separation controls alignment; emerged modes increase approximation error or unsupported mass.

For aligned coordinates, drift decomposes exactly as

$$G^+ - G^- = D^-(b^+ - b^-) + (D^+ - D^-)b^+. \tag{9}$$

The first term is projected incidence movement in the pre-change response geometry; the second is efficacy movement evaluated on the post-change mixture. This predicts that belief-only adaptation should suffice for pure incidence drift, response-only adaptation for pure efficacy drift, and both are needed when the two move together.

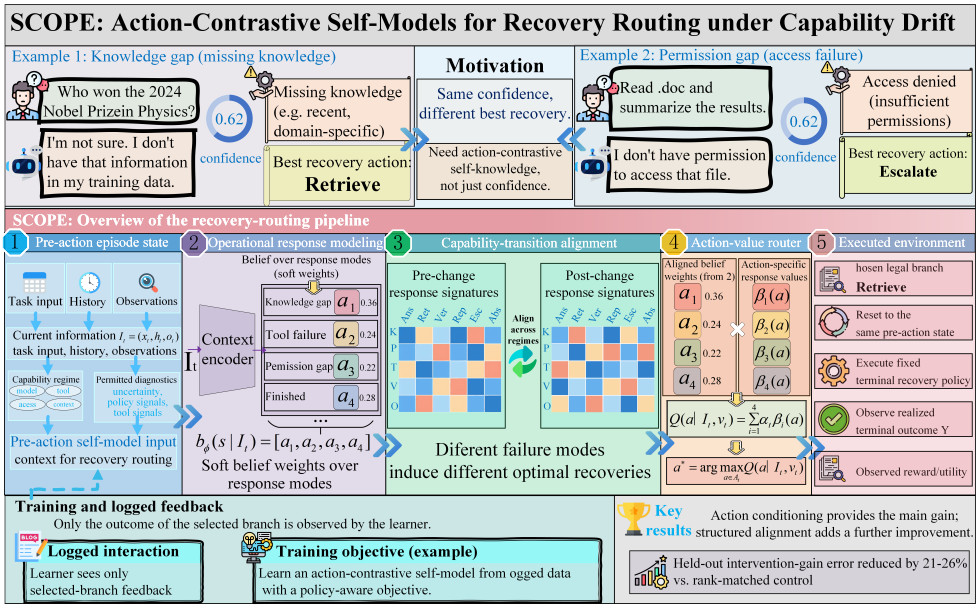


Figure 2: SCOPE recovery routing. The displayed modes, weights, and signatures are schematic; $M(a, i)$ abbreviates $M_{v_t}(a, i)$. It aligns persistent response modes, predicts action contrasts $\hat{G} = \hat{D}\hat{b}$, executes the highest-net-value legal branch, and learns response values from selected-branch feedback.

4 SCOPE

Figure 2 summarizes the information flow. SCOPE learns a soft belief encoder $\hat{b}_\theta(I, v)$, regime-conditioned response matrices $\hat{M}_\omega(v)$, and diagnostic emissions $\hat{L}_\psi(v)$. It never requires a hard semantic cause label. The learned modes are operational equivalence classes: they are useful only insofar as they support distinct branch-outcome predictions.

Learning uses only the realized outcome of the selected branch. If branch a_t was sampled by a logging policy with known propensity $p_t(a_t \mid I_t, v_t)$, the outcome component is

$$\mathcal{L}_{\text{out}} = \frac{1}{n} \sum_{t=1}^{n} \frac{\ell\left(e_{a_t}^\top \hat{M}_{v_t} \hat{b}(I_t, v_t), Y_t^{a_t}\right)}{p_t(a_t \mid I_t, v_t)}. \tag{10}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{out}} + \beta \mathcal{L}_{\text{diag}} + \gamma \mathcal{R}, \tag{11}$$

where the diagnostic term encourages modes to preserve observable response structure and $\mathcal{R}$ bounds factor norms and discourages context-specific identifier modes. Diagnostics provide an auxiliary predictive signal; they do not assign semantic names.

After a transition, SCOPE estimates standardized signatures on the common supported actions, solves a thresholded partial minimum-cost assignment to persistent signatures, and sends matches beyond the rejection threshold to an unknown component. Matched columns are warm-started, while unsupported cells retain conservative estimates until selected outcomes arrive. The router then computes Eq. 1 from $\hat{M}\hat{b}$ and executes the highest-valued legal branch.

Appendix D specifies the update and routing steps.

Direct-AV receives the same context, diagnostics, legal-action mask, propensities, selected outcomes, and costs, but predicts each action value without factorization. LowRank-AV uses the same K, shared encoder class, and comparable parameter scale as SCOPE, but its factors receive no diagnostic loss, cross-version signature matching, or unknown-state rejection. Direct-AV therefore tests whether sharing is useful; LowRank-AV tests what the diagnostic objective, alignment, and rejection add beyond generic low-rank regularization.

This comparison also states what SCOPE cannot infer. Without action support, neither a factorized nor a direct predictor identifies an untried branch. SCOPE can transport a response pattern only when the signature separation in Proposition 2 remains credible.

5 Experimental Design

Real executions and the paired unit. All terminal outcomes are produced by actual local-model, retrieval, verifier, container, and tool executions. The basic unit is an item–checkpoint–branch condition. For each source item and checkpoint, we freeze the pre-action prompt, environment snapshot, legal-action mask, retrieval index, tool state, and evaluator version, then execute every legal recovery branch from that identical state. Each branch condition is repeated with decoding seeds 17, 29, and 41. Parse failures, timeouts, invalid tool calls, and policy violations remain in the denominator. The evaluation set contains 1,800 source items, 10,800 model–item episodes, 58,080 distinct branch conditions, and 174,240 realized branch executions.

Learner-facing feedback versus evaluator-facing rollouts. Complete branch execution is used only to build a paired evaluator. For each model–item–repeat tuple, one legal branch is sampled by the logged policy and the learner receives $(I_t, v_t, a_t, p_t, Y_t^{a_t})$. Alternative branch outcomes, evaluator fields, latent-mode annotations, and future diagnostics are hidden. The held-out executed branches provide same-item oracle regret and supported value error, but never enter policy fitting. The primary log samples uniformly over legal branches; a skewed low-overlap logger is reserved for the support stress test.

Models and serving. We evaluate six open-weight checkpoints spanning dense, sparse-MoE, instruction, and reasoning-oriented model families: Qwen3.6-27B, Qwen3.6-35B-A3B, Gemma-4-E4B-it, Ministral-3-14B-Instruct-2512, MiMo-7B-RL, and Phi-4-mini-instruct (Gemma Team, 2026; Liu et al., 2026; LLM-Core Xiaomi, 2025; Microsoft et al., 2025). All rollouts run on one 8×NVIDIA A100-SXM4-80GB node through vLLM 0.25.1 (Kwon et al., 2023) with BF16 weights and the official chat template. Models are served sequentially to avoid cross-checkpoint contention. We cap context at 32,768 tokens, visible generation at 1,024 tokens (2,048 for repair/replanning), and use temperature 0.2 with top-p 0.95. Table 4 records the serving and parallel layout.

Tasks, branches, and repair blindness. The 1,800 items cover knowledge (360), checker-scored mathematics (360), stateful tool use (300), permission/access changes (300), context removal or compression (240), and repository repair (240). Task carriers use MMLU, MATH, ToolSandbox, and SWE-bench-style repository issues (Hendrycks et al., 2021a;b; Lu et al., 2025; Jimenez et al., 2024). Legal branches are answer, abstain, retrieve-then-answer, verify-then-revise, replan-then-answer, and escalate to a fixed fallback. In the code subset, every branch begins from the same clean checkout. The repair policy may inspect the issue, repository, and public tests exposed in the prompt, but never hidden tests, reference patches, alternative-branch outputs, or evaluator feedback. Hidden tests run once after the final patch; their failures cannot trigger an in-episode retry. We call this the repair-blind protocol.

Transitions and controlled structure tests. Executed transitions replace one component and rerun the same items: the base checkpoint, verifier, retrieval index, tool backend, permission mask, or supplied context. Stable reruns estimate ordinary execution variance. Incidence-only analyses reweight observed rollout strata while keeping their measured branch responses; efficacy-only analyses hold the item mixture fixed while substituting an executed component; joint transitions change both. Permutation changes only the mode correspondence supplied to alignment. Merge and split combine or partition observed response strata. New-state tests withhold one real failure family until post-change, and diagnostic-noise tests corrupt only the observable diagnostic channel. No terminal success value is generated analytically in these interventions.

Splits, repeats, and uncertainty. Source items are split 60/20/20 into 1,080 train, 360 validation, and 360 test items before prompt variants or branch runs are created. Every branch has 3 decoding repeats; SCOPE, LowRank-AV, and Direct-AV are refit with training seeds 3, 11, 29, 47, and 83. Hyperparameters are selected once on validation transitions and frozen for leave-one-checkpoint,

Table 1: Aggregate results across 6 checkpoints. Intervals are paired source-item bootstrap intervals.

Method	Utility [95% CI] ↑	Regret ↓	Brier ↓	Transfer gap ↓	Cost ↓
Answer-Only	.638 [.635,.641]	.151	–	–	.000
Scalar	.681 [.678,.684]	.108	.187	.075	.025
History	.694 [.691,.697]	.095	.172	.057	.030
StationaryVoI	.704 [.701,.707]	.085	.160	.050	.042
Direct-AV	.718 [.715,.721]	.071	.148	.039	.046
LowRank-AV	.725 [.722,.728]	.064	.141	.032	.047
SCOPE	**.731 [.728,.734]**	**.058**	**.134**	**.024**	.049
ActionOracle	.789 [.786,.792]	.000	–	–	.052

leave-one-transition, and leave-one-task–action experiments. We report the mean over training seeds and use 10,000 paired cluster-bootstrap resamples over source items, keeping checkpoints, branches, and decoding repeats for an item in the same cluster.

Comparators and metrics. Answer-Only, Scalar, History, ToolHealth, and StationaryVoI provide increasingly informed but transition-limited policies. Direct-AV is the matched unrestricted outcome predictor; LowRank-AV has the same shared rank and inputs as SCOPE but no diagnostic objective, signature matching, or unknown-state rejection. BeliefOnly and ResponseOnly isolate the two terms in Eq. 9; ActionOracle is an evaluator ceiling. Primary metrics are normalized net utility and same-item oracle regret. We additionally report supported action-value Brier score, intervention-gain MAE, transfer gap, alignment accuracy, unknown-state detection, branch cost, and training-seed variation.

6 Results

6.1 Action Conditioning Is the Main Gain; Structure Adds a Smaller One

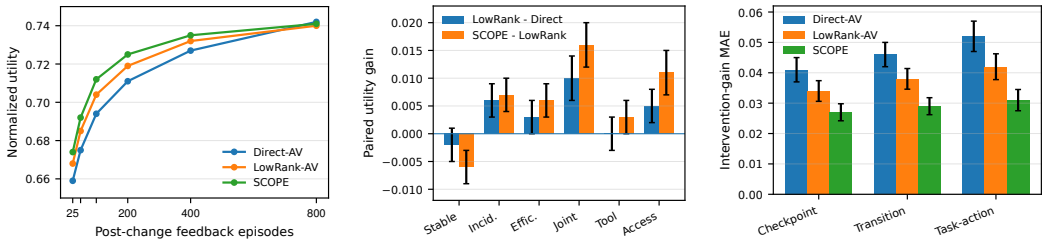
Table 1 separates the three steps that a Scalar-to-SCOPE comparison would conflate. Direct action prediction raises utility from .681 to .718. Generic shared-rank prediction reaches .725, and SCOPE reaches .731. Direct paired intervals are Direct-AV$-$Scalar $= +.037$ [+.034,+.040], LowRank-AV$-$Direct-AV $= +.007$ [+.005,+.009], and SCOPE$-$LowRank-AV $= +.006$ [+.004,+.008]. Thus the paper's largest empirical claim is action conditioning; the SCOPE-specific claim is the smaller increment of diagnostic prediction, signature alignment, and unknown-mode rejection together.

The SCOPE–LowRank increment is .006, .007, .006, .006, .006, and .005 for Qwen3.6-27B, Qwen3.6-35B-A3B, Gemma-4-E4B-it, Ministral-3-14B, MiMo-7B-RL, and Phi-4-mini, respectively. Across 5 refits, the mean utility standard deviations are .0034 for Direct-AV, .0028 for LowRank-AV, and .0026 for SCOPE. These summarize refit variability; the paired source-item intervals quantify a different source of uncertainty.

6.2 The Gain Appears Where Sharing Should Matter

Figure 3(a) varies post-change feedback. At 25 episodes, SCOPE reaches .674 versus .668 for LowRank-AV and .659 for Direct-AV. Direct-AV catches up by 800 episodes (.742 versus .741), matching the predicted variance–bias crossover. The drift decomposition is similarly selective (Figure 3(b)): SCOPE trails in stable reruns, adds .007 over LowRank-AV under incidence drift, .006 under efficacy drift, .016 under joint drift, and .011 under access/context change. BeliefOnly is strongest among one-component ablations for incidence movement; ResponseOnly is strongest for efficacy movement.

Held-out gain prediction provides a numerical transfer test rather than a post-hoc policy ranking. SCOPE yields MAE .027, .029, and .031 when a checkpoint, transition family, or task–action composition is excluded during fitting, compared with .034, .038, and .042 for LowRank-AV and .041, .046, and .052 for Direct-AV (Figure 3(c)). Table 2 additionally shows that R^2 improves across all three splits: from .54 to .66, .47 to .61, and .41 to .58 over LowRank-AV. The relative MAE re-



(a) Feedback budget (b) Drift structure (c) Held-out prediction

Figure 3: Feedback and transfer. (a) Sharing helps with scarce post-change feedback, while Direct-AV catches up at 800 episodes. (b) Joint incidence and efficacy drift yields the largest SCOPE increment over LowRank-AV. (c) SCOPE reduces held-out gain error. Where shown in (b,c), error bars are paired 95% source-item bootstrap intervals.

Table 2: Prediction of intervention gain in settings excluded during fitting.

Held out	Gain MAE			R^2		
	Direct	LowRank	SCOPE	Direct	LowRank	SCOPE
Model	.041	.034	**.027**	.42	.54	**.66**
Drift	.046	.038	**.029**	.35	.47	**.61**
Task–action	.052	.042	**.031**	.28	.41	**.58**

ductions are 20.6%, 23.7%, and 26.2%, respectively. These are out-of-split prediction comparisons, rather than merely a favorable ranking of final policies.

6.3 Alignment Helps Only While the Response Geometry Persists

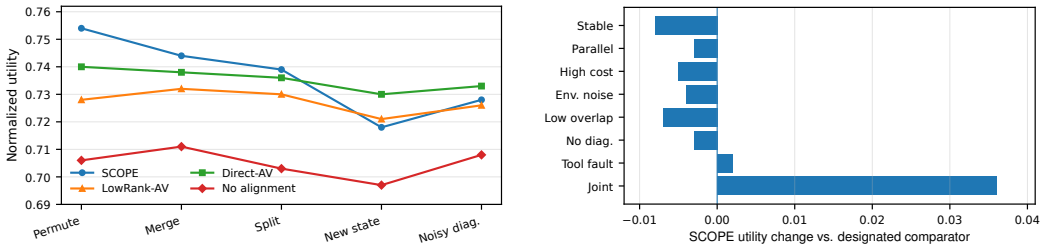
Figure 4(a) tests correspondence under structural change. Under a pure permutation, signature matching attains .93 persistent-mode correspondence accuracy and .754 utility, compared with .728 for LowRank-AV, .740 for Direct-AV, and .706 without alignment. Merge and split narrow the margin. The reversal is more informative: on a withheld new response family, SCOPE reaches .718 and detects 79% of unmatched cases, while Direct-AV reaches .730; under noisy diagnostics, Direct-AV again leads .733 to .728. The no-alignment ablation performs worse in both cases. These failures make Proposition 2 a testable condition rather than a semantic-label claim.

Figure 4(b) extends the comparison to decision boundaries. When intervention cost exceeds the recoverable success mass, a simpler abstention or confidence rule suffices. Low overlap leaves changed response cells weakly identified, and transient environmental noise need not support a persistent response mode. In contrast, a directly observed tool fault is largely captured by tool health alone. The joint-drift condition is favorable because both mode incidence and branch efficacy must adapt; the benefit is conditional on that reusable structure.

6.4 Three Executed Cases Explain the Aggregate

Equal confidence, opposite recovery. Matched knowledge-gap and permission-gap items both have direct-answer success near .62. Scalar routing chooses the correct recovery on .50 of pairs; Direct-AV reaches .71, SCOPE .82, and the executed oracle .91. Retrieval and escalation exchange rank while confidence remains fixed.

Stable answer accuracy, changing verifier efficacy. After a verifier replacement, direct-answer success remains .72, but the fraction of previously failed items rescued by verification falls from .75 to .41. BeliefOnly sees little incidence movement and continues verifying. SCOPE updates the verification response column and shifts toward replanning; Direct-AV reaches the same choice only after more selected verification outcomes.



(a) Alignment and structural change

(b) Boundaries of useful sharing

Figure 4: When structure transfers and when it fails. (a) Signature alignment helps under permutation; new modes and corrupted diagnostics favor Direct-AV. (b) Utility differences against the designated comparator for each condition show that stable behavior, high costs, and weak overlap can erase the benefit.

A new repair failure breaks the old basis. A repository/runtime transition introduces patches that pass public tests but fail a new hidden integration constraint. Under the repair-blind protocol, the failure is not visible before submission and its action signature is absent pre-change. SCOPE's unknown component rejects most forced matches but misses some, so Direct-AV wins the new-state subset. The boundary is structural, not repaired by renaming the mode.

7 Discussion and Conclusion

Action-conditioned prediction accounts for the largest gain; generic sharing contributes a smaller low-feedback benefit, and aligned diagnostic modes add value when their response geometry persists. SCOPE therefore provides a conditional model of recovery, with action support, shared-basis fit, and signature separation determining when it is useful.

Routing depends on decision margins as well as prediction error. If all legal action values are estimated within ε, the greedy action incurs at most 2ε regret; a best-action margin greater than 2ε preserves the selected action. Thus a lower average prediction error need not yield higher utility when the remaining errors occur near cost-sensitive decision boundaries. Appendix B gives the argument.

The evaluation concerns one-step selection among fixed terminal recovery policies. Paired reset executions support within-item comparisons, while the learner observes only one selected outcome. Incidence-only analyses reweight observed strata and do not reproduce every natural deployment shift. New modes, merges, corrupted diagnostics, and weak support remain failure conditions. Extending the framework to learned multi-step recovery requires a sequential state and policy model. Within the present scope, the central distinction is operational: confidence predicts whether an agent may fail; recovery knowledge predicts how each available action changes terminal value.

AI USE STATEMENT

We used LLMs solely as general-purpose writing assistants in preparing this submission. Their use was limited to improving the clarity, readability, and conciseness of the manuscript, including refining sentence structure, correcting grammar, and reducing unnecessary verbosity. All research ideas, methodological development, experiments, analyses, interpretations, and final writing decisions were independently conceived, conducted, and verified by the authors.

REPRODUCIBILITY STATEMENT

Section 5 describes the paired execution and selected-feedback protocol. Appendix B states the mathematical assumptions and proofs; Appendices C and D describe learning and routing; Appendix F specifies sample accounting, information boundaries, and paired evaluation.

REFERENCES

Aldo Gael Carranza, Sanath Kumar Krishnamurthy, and Susan Athey. Flexible and Efficient Contextual Bandits with Heterogeneous Treatment Effect Oracles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7190–7212, 2023. URL https://proceedings.mlr.press/v206/carranza23a.html.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-Level Uncertainty in Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=11QZITAMUO.

Lingjiao Chen, Matei Zaharia, and James Zou. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2), 2024. doi: 10.1162/99608f92.5317da47. URL https://hdsr.mitpress.mit.edu/pub/y95zitmz/release/2.

Mingwei Deng, Ville Kyrki, and Dominik Baumann. Transfer Learning in Latent Contextual Bandits with Covariate Shift through Causal Transportability. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 731–756, 2025. URL https://proceedings.mlr.press/v275/deng25a.html.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Miroslav Dudík, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A Model Update Strategy for Compatible LLM Evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7320–7332, 2024. doi: 10.18653/v1/2024.findings-emnlp.430. URL https://aclanthology.org/2024.findings-emnlp.430/.

Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Abinitha Gourabathina, Inkit Padhi, Manish Nagireddy, Subhajit Chaudhury, and Prasanna Sattigeri. Answering the Wrong Question: Reasoning Trace Inversion for Abstention in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13307–13324, 2026. doi: 10.18653/v1/2026.acl-long.608. URL https://aclanthology.org/2026.acl-long.608/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Joey Hong, Branislav Kveton, Manzil Zaheer, Sumeet Katariya, and Mohammad Ghavamzadeh. Multi-Task Off-Policy Learning from Bandit Feedback. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13157–13173, 2023. URL https://proceedings.mlr.press/v202/hong23a.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Nathan Kallus and Angela Zhou. Stateful Offline Contextual Policy Evaluation and Learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 11169–11194, 2022. URL https://proceedings.mlr.press/v151/kallus22a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective Generation for Controllable Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50494–50527, 2024. doi: 10.52202/079017-1600. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a6815122f533193a022cbc41786c1cc-Abstract-Conference.html.

Jiabin Lin, Shana Moothedath, and Namrata Vaswani. Fast and Sample Efficient Multi-Task Representation Learning in Stochastic Contextual Bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30227–30251, 2024. URL https://proceedings.mlr.press/v235/lin24o.html.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive Representations of State. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Alexander H. Liu et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026. URL https://arxiv.org/abs/2601.08584.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. Uncertainty Calibration for Tool-Using Language Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16781–16805, 2024a. doi: 10.18653/v1/2024.findings-emnlp.978. URL https://aclanthology.org/2024.findings-emnlp.978/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=zAdUB0aCTQ.

LLM-Core Xiaomi. MiMo: Unlocking the Reasoning Potential of Language Model – From Pretraining to Posttraining. *arXiv preprint arXiv:2505.07608*, 2025. URL https://arxiv.org/abs/2505.07608.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping 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*, pp. 1160–1183, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Microsoft, Abdelrahman Abouelenin, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Elliot Nelson, Debarun Bhattacharjya, Tian Gao, Miao Liu, Djallel Bouneffouf, and Pascal Poupart. Linearizing Contextual Bandits with Latent State Dynamics. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1477–1487, 2022. URL https://proceedings.mlr.press/v180/nelson22a.html.

Adith Swaminathan and Thorsten Joachims. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Chris van Merwijk, Ryan Carey, and Tom Everitt. A Complete Criterion for Value of Information in Soluble Influence Diagrams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10034–10041, 2022. doi: 10.1609/aaai.v36i9.21242. URL https://ojs.aaai.org/index.php/AAAI/article/view/21242.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and Adaptive Off-policy Evaluation in Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3589–3597, 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

Zhangyi Wang, Jiexiang Xu, Bingnan Yu, and Zongze Li. ARCHITECT: Uncertainty-Aware Dynamic Tool Learning via Causal Intervention for Open-World Agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36819–36829, 2026. doi: 10.18653/v1/2026.acl-long.1707. URL https://aclanthology.org/2026.acl-long.1707/.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Yunze Xiao, Junjue Wang, and Naoto Yokoya. The Confidence Dichotomy: Analyzing and Mitigating Miscalibration in Tool-Use Agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11325–11349, 2026. doi: 10.18653/v1/2026.acl-long.520. URL https://aclanthology.org/2026.acl-long.520/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

APPENDIX

A NOTATION AND SCOPE

An episode provides pre-action information I, regime v, and legal branch set $\mathcal{A}_v$. A branch a is a fixed terminal policy such as answering, retrieving and answering, verifying and answering, replanning, escalating, or abstaining. Its same-world terminal success is $Y^a \in \{0, 1\}$. We evaluate

$$V_v(a \mid I) = \mathbb{E}[Y^a \mid I, v] + u_v(a) - \lambda k_v(a), \tag{12}$$

where u_v contains non-success utility and k_v is branch cost.

The learned response variable S has K soft modes, including an `ok` mode. It is not a provided failure label. SCOPE learns

$$b_v(s \mid I) = P(S = s \mid I, v), \tag{13}$$

$$M_v(a, s) = \mathbb{E}[Y^a \mid S = s, v], \tag{14}$$

$$D_v(a, s) = M_v(a, s) - M_v(a_0, s), \tag{15}$$

where a_0 is the answer branch. The central modeling condition is conditional response homogeneity:

$$\mathbb{E}[Y^a \mid S = s, I, v] = M_v(a, s). \tag{16}$$

This condition says that the mode retains the part of context that predicts how branches respond. It need not retain every semantic distinction. If two semantic causes induce the same action effects they may share a mode; if one cause has two different rescue patterns it should split.

B PROOFS AND COUNTEREXAMPLES

B.1 THE EXACT BOUNDARY FOR SCALAR SUFFICIENCY

Fix a capability regime and write $m_0 = M(a_0, :)^\top$, so that $q(b) = m_0^\top b$ and $G(b) = Db$ for $b \in \Delta^{K-1}$. Assume that all mixtures in the simplex are admissible and that branch offsets may vary over a family rich enough to separate any two distinct action contrasts. Then confidence q preserves every branch preference if and only if

$$\ker \begin{bmatrix} \mathbf{1}^\top \\ m_0^\top \end{bmatrix} \subseteq \ker D. \tag{17}$$

Equivalently, there exist vectors $\alpha, \beta \in \mathbb{R}^A$ such that

$$D = \alpha \mathbf{1}^\top + \beta m_0^\top, \qquad G(b) = \alpha + \beta q(b). \tag{18}$$

Thus mode-invariant action effects are sufficient but are not necessary: a mode-dependent contrast is still recoverable from confidence when it is an affine function of that mode's answer-success probability.

Proof. If Eq. 18 holds, Db depends on b only through $q(b)$. Conversely, let z satisfy $\mathbf{1}^\top z = m_0^\top z = 0$ and let b_* be an interior simplex point. For sufficiently small $t > 0$, both $b_* + tz$ and $b_* - tz$ are valid mixtures with equal confidence. Sufficiency over the rich offset family forces their contrasts to coincide, hence $Dz = 0$. The row-space/kernel duality then gives Eq. 18. For a restricted set of attainable mixtures, the kernel condition need only hold on differences between attainable equal-confidence beliefs. □

The rescue-only special case. Suppose a recovery never damages a case that the answer branch would solve. Let $R(a, c)$ be its success probability conditional on answer failure of type c, let $R(a_0, c) = 0$, and let $0 < q < 1$. Then

$$V(a) - V(a') = (1 - q) \sum_c \pi(c)[R(a, c) - R(a', c)] + u(a) - u(a') - \lambda[k(a) - k(a')]. \tag{19}$$

For fixed q and arbitrary failure mixtures, confidence is sufficient exactly when each pairwise rescue difference is constant across failure types. If the difference varies between two types, point masses on those types give equal confidence but different contrasts; an intermediate offset reverses their ordering. The assumption excluding damage is essential for this rescue-only formula. General recovery policies are covered by Eqs. 17–18.

Illustrative calculation. For $q = 0.62$, consider retrieval rescue rates $(0.82, 0.08)$ and escalation rescue rates $(0.07, 0.84)$ for knowledge and permission failures. Under the no-damage assumption, success equals $q + (1 - q)R(a, c)$:

Failure type	Retrieval success	Escalation success
Knowledge	0.9316	0.6466
Permission	0.6504	0.9392

These numbers are an algebraic illustration, not additional measured outcomes. Equal costs preserve the reversal; a sufficiently large common intervention cost makes answering preferable in both cases.

B.2 Decision Sufficiency and Minimal Coordinates

Under conditional response homogeneity, subtracting the reference branch gives

$$V(a \mid I, v) - V(a_0 \mid I, v) = [D_v b_v(I)]_a + u_v(a) - u_v(a_0) - \lambda[k_v(a) - k_v(a_0)]. \tag{20}$$

Consequently $(D_v b_v, u_v, k_v, \mathcal{A}_v)$ suffices for branch choice. The reference success probability is needed only to recover absolute values. For fixed D_v, the equivalence classes of beliefs are exactly the fibers of $b \mapsto D_v b$. Let H have orthonormal columns spanning $\{z : \mathbf{1}^\top z = 0\}$. On the full simplex, the minimal number of linear contrast coordinates is $\operatorname{rank}(D_v H)$; $D_v b$ itself may contain redundant coordinates. Any statistic preserving all preferences over a rich offset family must distinguish any two such fibers.

B.3 A Multi-Regime Estimation Bound

Let $Z = (I, v)$, $v \in \{1, \ldots, R\}$, and assume independent logged samples (Z_i, A_i, Y_i) with $P(A_i = a \mid Z_i) = p_i(a \mid Z_i) \geq p_{\min}$ on legal actions. Assume consistency and that action selection is conditionally independent of the potential outcomes given Z_i. Define the average legal-action risk

$$\mathcal{R}(f) = \mathbb{E}_Z \left[\frac{1}{|\mathcal{A}(Z)|} \sum_{a \in \mathcal{A}(Z)} \mathbb{E}[(f(Z, a) - Y^a)^2 \mid Z] \right]. \tag{21}$$

For predictions in $[0, 1]$, its unbiased estimator is

$$\widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{(f(Z_i, A_i) - Y_i)^2}{|\mathcal{A}(Z_i)| p_i(A_i \mid Z_i)}. \tag{22}$$

Omitting $|\mathcal{A}(Z_i)|$ instead targets the sum over legal actions, which weights contexts with more available branches more heavily.

For $\|\phi(I)\|_2 \leq 1$, consider bounded classes $f(Z, a) = e_a^\top W_v \phi(I)$ and $f(Z, a) = e_a^\top M_v B \phi(I)$ with a *single shared* matrix $B \in \mathbb{R}^{K \times d}$. Their parameter-count cover bounds use

$$p_D = RAd, \qquad p_S = Kd + RAK. \tag{23}$$

Here A is the size of the action vocabulary; legality is imposed by masking. Bounded parameter norms and a Lipschitz parameter-to-prediction map give log-cover size at most $Cp \log(C'/\epsilon)$ for either class. Each weighted loss is at most $1/p_{\min}$, and its conditional second moment is at most $1/p_{\min}$ because

$$\sum_{a \in \mathcal{A}(Z)} p(a \mid Z) \left(\frac{\ell_a}{|\mathcal{A}(Z)| p(a \mid Z)} \right)^2 \leq \frac{1}{|\mathcal{A}(Z)| p_{\min}} \leq \frac{1}{p_{\min}}. \tag{24}$$

A Bernstein bound, a union bound over the cover, and Lipschitz extension therefore yield, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\widehat{\mathcal{R}}(f) - \mathcal{R}(f)| \leq C_1 \sqrt{\frac{p \log(C_2 n) + \log(2/\delta)}{n p_{\min}}} + C_3 \frac{p \log(C_2 n) + \log(2/\delta)}{n p_{\min}}. \tag{25}$$

If $\hat{f}$ minimizes empirical risk to optimization error $\varepsilon_{\mathrm{opt}}$, comparison with a population minimizer gives

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \le 2 \sup_{f \in \mathcal{F}} |\widehat{\mathcal{R}}(f) - \mathcal{R}(f)| + \varepsilon_{\mathrm{opt}}. \tag{26}$$

Misspecification adds $\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}(f_{\mathrm{Bayes}})$. The square-root terms quoted in the main text are the leading terms in the regime $np_{\min}$ larger than the effective log-cover dimension. For a fixed unnormalized sum-risk with at most A actions, the variance bound acquires an A factor. Correlated repetitions of the same source item require a cluster-level concentration argument; raw rollout count is not automatically the number of independent samples in Eq. 25.

The dimension comparison applies to LowRank-AV as well as SCOPE and requires $K(d + RA) < RAd$. It neither proves that the fitted nonlinear encoder belongs to this linear class nor establishes a statistical advantage for a regime-specific encoder B_v. With $K = A = 6$, factorization need not reduce the capacity within one regime; the displayed saving can arise from sharing a common basis across regimes. Latent coordinates themselves are not identified by this parameter counting argument.

B.4 Alignment with Emerged or Disappeared Modes

Compare signatures only on common supported action coordinates and on diagnostic coordinates expressed in the same fixed units. Suppose each persistent old mode s has a unique persistent new partner $\pi(s)$ with true distance at most η. Suppose all incorrect pairs, including pairs involving new or vanished modes, have distance at least Δ, and every signature estimate has norm error at most ϵ. If $\Delta > \eta + 4\epsilon$, choose

$$\eta + 2\epsilon < \tau < \Delta - 2\epsilon. \tag{27}$$

Construct the bipartite graph containing only edges whose estimated distance is at most τ, and compute a maximum-cardinality, minimum-cost *partial* matching. Correct edges have estimated length at most $\eta + 2\epsilon$, while incorrect edges have length at least $\Delta - 2\epsilon$. The thresholded graph therefore consists exactly of the disjoint persistent pairs. Its matching recovers those pairs and leaves new or disappeared modes unmatched.

A complete assignment that forces all modes to match does not have this rejection guarantee. Likewise, the result assumes consistently estimated mode-specific signatures; response homogeneity alone does not identify a latent factorization up to permutation. A lower distance threshold rejects more matches, while a higher threshold admits less certain correspondence.

B.5 From Value Error to One-Step Regret

Let $a^* \in \arg\max_a V(a)$ and $\hat{a} \in \arg\max_a \hat{V}(a)$ over the same legal set. On an event where $\max_a |\hat{V}(a) - V(a)| \le \xi$,

$$V(a^*) - V(\hat{a}) = [V(a^*) - \hat{V}(a^*)] + [\hat{V}(a^*) - \hat{V}(\hat{a})] + [\hat{V}(\hat{a}) - V(\hat{a})] \tag{28}$$

$$\le 2\xi. \tag{29}$$

If the event fails with probability at most ρ and legal values span at most B, then $\mathbb{E}[\mathrm{Regret}] \le 2\xi + B\rho$. Taking ξ as the sum of estimation, alignment, and approximation errors gives the accounting bound in the main paper. The average squared-risk bound above does not by itself provide the required uniform value bound; converting between them needs an additional coverage or margin argument.

B.6 Incidence and Efficacy Drift

For aligned coordinates, adding and subtracting $D^- b^+$ gives

$$D^+ b^+ - D^- b^- = (D^+ b^+ - D^- b^+) + (D^- b^+ - D^- b^-) \tag{30}$$

$$= (D^+ - D^-) b^+ + D^- (b^+ - b^-). \tag{31}$$

The first term changes branch efficacy on the post-change mixture, while the second changes mode incidence in the pre-change geometry. This identity has an operational mode interpretation only after correspondence has been established. A new mode or a split can invalidate that interpretation without preventing direct prediction of branch values.

High-cost boundary. If $q = .80$, no recovery damages correct answers, and every paid branch has net cost .30, then even a perfect rescue supplies at most $1 - q = .20$ additional success. Every such intervention loses to answering by at least .10. Distinct response signatures need not change the selected action at a fixed prohibitive cost.

C Learning and Alignment Details

SCOPE jointly learns a soft encoder $\hat{b}_\theta(I, v)$, response matrix $\hat{M}_\omega(v)$, and diagnostic emissions $\hat{L}_\psi(v)$. The training loss is the selected-branch outcome loss plus diagnostic prediction and bounded regularization terms. Logged propensities weight selected outcomes. Unchosen potential outcomes are reserved for evaluation and never enter the policy update.

At a version change, SCOPE builds signatures from supported response columns and diagnostics, solves a thresholded partial minimum-cost assignment, and sends signatures above the rejection threshold to an unknown component. Matched columns are warm-started; unmatched columns retain conservative values until supported outcomes arrive. The branch selector then evaluates Eq. 12 over legal actions.

Transition update. (1) Encode each task as a soft response distribution. (2) Estimate action-response and diagnostic signatures from supported outcomes. (3) match persistent modes and reject unsupported matches. (4) Update matched and unknown components with selected-action feedback. (5) choose the legal action with largest predicted net value.

Direct-AV uses the same context representation, diagnostic fields, legal-action mask, logged propensities, selected outcomes, and cost function. It replaces $\hat{M}\hat{b}$ with independent action-value heads. LowRank-AV keeps the rank and parameter scale of $\hat{M}\hat{b}$ but removes the diagnostic loss, signature matching, and unknown-state mechanism. Direct-AV therefore tests generic sharing; LowRank-AV tests the additional operational structure specific to SCOPE.

C.1 Implementation and Model Selection

Table 3 records the configuration fixed before held-out evaluation. Validation maximizes normalized utility subject to the common action coverage constraint; no held-out model, drift family, or task–action composition enters selection.

Table 3: Implementation settings used in the reported study.

Component	Frozen setting
Belief encoder	Two-layer MLP, width 256, GELU
Response modes	$K = 6$ including the `ok` component
Optimizer	AdamW, learning rate 3×10^{-4}
Training	40 epochs; patience 6 on validation utility
Outcome / diagnostic weights	$1.0/0.25$
Factor regularization	0.01 Frobenius and entropy penalty
Signature metric	Standardized ℓ_2 distance on $[M(:, s), L(:, s)]$
Unknown-state threshold	0.35
Logging coverage floor	$p_{\min} = 0.10$
Cost weight	$\lambda = 0.50$

C.2 Identifiability from Selected Actions

Chosen-action feedback identifies a response entry only under positive support, sufficient post-change observations, and either usable state attribution or a validated sharing model. To see the limitation, consider actions A and B and a policy that always chooses A. Worlds with $(V(A), V(B)) = (0.7, 0.4)$ and $(0.7, 0.9)$ induce identical logs but opposite optimal actions. Neither SCOPE, LowRank-AV, nor Direct-AV can identify B without exploration or valid transport. This motivates the shared coverage floor and the low-overlap failure condition rather than an assumption of complete counterfactual labels.

D Detailed Algorithmic Specification

This section specifies the statistical estimator, transition update, and information available to each operation. It does not prescribe the internal prompts of the terminal recovery policies. Each branch is held fixed during a routing comparison, including its tools, stopping rule, and terminal scorer.

D.1 Objects and Observable Information

For each logged episode, the policy record is

$$\mathcal{Z}_t = (I_t, v_t, \mathcal{A}_t, a_t, p_t, Y_t^{a_t}, z_t, r_t, u_t, k_t), \tag{32}$$

where I_t is the pre-action context, v_t is the regime, $\mathcal{A}_t$ is the legal branch set, p_t is the known probability of the selected branch, z_t contains permitted diagnostics available before action selection, and r_t marks observed diagnostic coordinates. The response of an unselected branch, hidden evaluator fields, and future diagnostics are not policy inputs. Diagnostics already contained in I_t need not be stored twice.

The encoder returns a probability vector $\hat{b}_\theta(I_t, v_t) \in \Delta^{K-1}$. The response parameters represent $\hat{M}_v(a, s) \in [0, 1]$, and the diagnostic parameters represent emissions $\hat{L}_v(:, s)$ in the diagnostic observation space. Branch success is predicted as

$$\hat{\mu}_t(a) = \sum_{s=1}^{K} \hat{M}_{v_t}(a, s)\hat{b}_\theta(s \mid I_t, v_t). \tag{33}$$

For a numerical emission parameterization, the corresponding diagnostic prediction is $\hat{z}_t = \hat{L}_{v_t}\hat{b}_\theta(I_t, v_t)$. Categorical or other diagnostic channels require their compatible emission loss; no particular diagnostic likelihood follows from the response-homogeneity model.

D.2 Selected-Outcome Training

Let $\mathcal{D}$ contain training records only. The outcome objective is

$$\widehat{\mathcal{L}}_{\text{out}} = \frac{1}{|\mathcal{D}|} \sum_{t \in \mathcal{D}} \frac{\ell_{\text{out}}(\hat{\mu}_t(a_t), Y_t^{a_t})}{p_t}. \tag{34}$$

This expression sums risk over the legal actions in expectation. In particular, it is distinct from an average-action objective when the number of legal branches varies. The propensity belongs to the logging decision that produced the record; it is not recomputed from the current router.

Write $\mathcal{J}_t = \{j : r_{tj} = 1\}$ for the observed diagnostic coordinates. An auxiliary emission objective has the form

$$\widehat{\mathcal{L}}_{\text{diag}} = \frac{1}{|\mathcal{D}_{\text{diag}}|} \sum_{t \in \mathcal{D}_{\text{diag}}} \ell_{\text{diag}}\big(\hat{z}_t[\mathcal{J}_t], z_t[\mathcal{J}_t]\big), \tag{35}$$

where $\mathcal{D}_{\text{diag}}$ contains records with observed diagnostic fields. This notation leaves the emission likelihood explicit as a model choice. A missing field contributes no diagnostic target. The complete fit sets this auxiliary term to zero if no diagnostic records are present and minimizes

$$\widehat{\mathcal{L}}_{\text{out}} + \beta\widehat{\mathcal{L}}_{\text{diag}} + \gamma\mathcal{R}(\theta, \{M_v, L_v\}_v), \tag{36}$$

with the selected validation settings. Joint fitting does not require a semantic failure label or a hard assignment of each item to a mode. Because the factors are coupled, a gradient update can change predictions for more than the selected action; such changes are model-based sharing, not observed outcomes for the unselected actions.

D.3 Transition Matching and Warm Starts

At a regime transition, preserve the previous response and diagnostic parameters. Estimate candidate post-change signatures using available selected-action records and permitted diagnostics. Let

$\mathcal{C}$ be the common supported action coordinates, and retain diagnostic coordinates with the same measurement semantics on both sides. In a common standardization, form

$$\hat{h}_v(s) = \text{standardize}\left([\hat{M}_v(\mathcal{C}, s), \hat{L}_v(:, s)]\right). \tag{37}$$

The same standardization must be used for both sides of a comparison; otherwise a small distance need not express a stable response signature.

For candidate old and new modes, set $C_{ss'} = \|\hat{h}_-(s) - \hat{h}_+(s')\|_2$. Retain only edges with $C_{ss'} \leq \tau$. Among these edges, maximize matching cardinality and then minimize total distance. This is a partial one-to-one assignment: a new or disappeared mode is allowed to remain unmatched. Accepted matches define the coordinate correspondence used to warm-start the current parameters from the previous response columns. Post-change selected feedback then updates their efficacy. Rejected candidates are tracked as unknown rather than being forced into an old identity.

The matching theorem does not specify a numerical initializer for unsupported response cells or a particular action to take on an unknown mode. Those choices belong to the implementation's conservative value policy and must be fixed independently of unobserved outcomes. Unknown-state rejection alone does not establish a bound on the value of an untried branch.

D.4 Complete Procedure

Algorithm: SCOPE with selected feedback and partial alignment

Inputs. Training logs with selected outcomes and known propensities; pre-action feature and diagnostic schemas; legal branch masks, utilities, and costs; a response-mode budget; diagnostic and regularization weights; a matching threshold; and a declared conservative policy for unsupported cells.

Maintained state. Encoder parameters, regime-specific response and diagnostic parameters, supported-coordinate information, and accepted or unknown mode correspondences.

1. **Fit the initial predictor.** For each training record, encode its pre-action context; compute branch and diagnostic predictions; evaluate the selected-outcome loss and observed diagnostic loss; and optimize their weighted sum with regularization.
2. **Detect the current regime.** Receive the model, tool, access, or context regime identifier. Preserve the previous parameters if the regime has changed. This step assumes the transition is supplied; it is not a change-point detection algorithm.
3. **Construct comparable signatures.** On a transition, use only currently available selected-action records and pre-action diagnostics to estimate candidate signatures on common supported coordinates.
4. **Match persistent coordinates.** Form threshold-eligible edges, solve a maximum-cardinality minimum-cost partial assignment, and warm-start accepted response columns from their matched predecessors. Mark rejected or unsupported candidates as unknown.
5. **Form current predictions.** Encode the current context and evaluate Eq. 33. Apply the declared conservative treatment to unsupported predictions; do not substitute evaluator-only outcomes or silently impute them as observed successes or failures.
6. **Select a legal branch.** Compute $\hat{V}_t(a) = \hat{\mu}_t(a) + u_t(a) - \lambda k_t(a)$ and choose $\hat{a}_t \in \arg\max_{a \in \mathcal{A}_t} \hat{V}_t(a)$ using a fixed tie-breaking rule. During randomized data collection, the actual branch is instead sampled from the declared logging policy, and its sampling probability is recorded.
7. **Execute and record.** Run the selected fixed terminal policy and observe its terminal outcome. Append the selected-action record with its historical propensity and available diagnostic mask.
8. **Update from admissible feedback.** Refit or update the shared predictor using the accumulated selected records and the same objective. Reassess support and, when a transition update is performed, the partial correspondence. Repeat for subsequent episodes.

Output. A selected legal recovery branch and an updated predictor. Unselected branch outcomes are never accessed by this procedure.

Randomized logging and greedy evaluation are different operations. A greedy deployment action does not supply positive propensity for every alternative; the identification assumptions apply to the data-collection policy that produced the available training records. Likewise, the algorithm specifies which records can be used, but does not determine the runtime feature extractor, minibatch schedule, signature-update frequency, precise diagnostic likelihood, or numerical unknown-state initializer. These details are needed for an executable reproduction and cannot be recovered from the mathematical factorization alone.

E Execution and Evaluation Details

Table 4: Local rollout inventory. TP and DP denote tensor- and data-parallel degree.

Checkpoint	Repository	Framework	Precision	TP	DP
Qwen3.6-27B	`Qwen/Qwen3.6-27B`	vLLM 0.25.1	BF16	4	2
Qwen3.6-35B-A3B	`Qwen/Qwen3.6-35B-A3B`	vLLM 0.25.1	BF16	4	2
Gemma-4-E4B-it	`google/gemma-4-e4b-it`	vLLM 0.25.1	BF16	1	8
Ministral-3-14B	`mistralai/Ministral-3-14B-Instruct-2512-BF16`	vLLM 0.25.1	BF16	2	4
MiMo-7B-RL	`XiaomiMiMo/MiMo-7B-RL`	vLLM 0.25.1	BF16	1	8
Phi-4-mini	`microsoft/Phi-4-mini-instruct`	vLLM 0.25.1	BF16	1	8

E.1 Branch Contracts and Information Boundaries

Each episode fixes the pre-action prompt, environment snapshot, legal-action mask, retrieval index, tool state, and evaluator version. A recovery branch is an option with a fixed terminal policy, rather than a single token or tool call. Table 5 defines the branch interface. The description specifies the information boundary; task-specific prompt wording and runtime configuration are part of the execution implementation.

Table 5: Terminal recovery interfaces. Availability is determined before routing.

Branch	Permitted operation	Terminal boundary
Answer	Respond from the current information.	Score the final answer or submitted artifact.
Retrieve	Query the fixed evidence source, then answer.	No alternative-branch output is visible.
Verify	Apply the designated verifier, then revise.	The verifier is distinct from the hidden evaluator.
Replan	Reconsider the solution within the fixed budget.	No hidden test feedback can trigger revision.
Escalate	Invoke the designated fallback policy.	Include fallback cost in branch value.
Abstain	Decline to complete the task.	Apply the task's non-success utility offset.

For repository repair, all branches start from the same clean checkout. The policy may read the issue, repository, and exposed public tests. Reference patches, hidden tests, alternative branch outputs, and evaluator feedback are excluded. Hidden tests evaluate the final patch once; failures cannot initiate an in-episode retry. Invalid calls, parse failures, timeouts, and policy violations count as unsuccessful outcomes.

E.2 Sample Accounting and Splits

The source-item counts are 360 knowledge, 360 mathematics, 300 stateful tools, 300 access, 240 context, and 240 repository-repair items, totaling 1,800. Across six checkpoints this gives 10,800 model–item episodes. The reported 58,080 legal branch conditions imply 5.378 legal branches per episode on average. Three decoding repeats give $58{,}080 \times 3 = 174{,}240$ branch executions. These

counts describe the base branch inventory; component-replacement reruns must be accounted for separately when measuring total execution cost.

The source-item split is 1,080/360/360 for training/validation/test, applied before variants and rollouts. Keeping every branch, repeat, and checkpoint of an item in its partition prevents the same source problem from entering both fitting and evaluation. Training seeds are 3, 11, 29, 47, and 83; decoding seeds are 17, 29, and 41. The evaluation unit for uncertainty is the source item, not an individual branch execution.

E.3 Logged Outcomes and Action Support

For each model–item–repeat tuple, the logger samples one legal action and records its known propensity. All methods receive the same pre-action context, diagnostics, legality mask, costs, selected action, propensity, and selected outcome. Other branch outcomes reside only in the evaluator. Uniform logging assigns $1/|\mathcal{A}_t|$ to each legal action; a propensity floor $p_{\min}$ is feasible only when $|\mathcal{A}_t|p_{\min} \leq 1$. A value of .10 is therefore feasible for the six-action vocabulary; .20 is not a feasible universal floor when all six branches are legal.

Availability and statistical support are different. An action may be legal but insufficiently sampled, while an illegal action is excluded from both routing and its target risk. Unknown-mode rejection handles doubtful transport; it does not identify an unobserved action outcome.

E.4 Paired Metrics and Bootstrap

For source item i, method j, checkpoint v, and decoding repeat r, let U_{ivrj} denote terminal utility on the common scale. Average over the prespecified checkpoint/repeat weights to obtain $\bar{U}_{ij}$, and compare methods through $d_i = \bar{U}_{ij} - \bar{U}_{ij'}$. Each bootstrap draw samples source-item indices with replacement and carries all associated methods, checkpoints, and repeats together. The reported intervals use 10,000 draws. Refits characterize training variation; branch executions from one item are not treated as independent clusters.

The action-value Brier score evaluates squared probability error on supported outcomes. Intervention gain is the difference between a branch and the answer reference. Oracle regret compares the selected branch with the best legal realized branch under the same evaluation convention. This realized oracle is an evaluation ceiling and can exceed the value of a policy optimizing conditional expected outcomes; it is not the Bayes-optimal router in Lemma 1.

F Experimental Protocol

F.1 Decision Cards

Each card fixes a task, legal branches, costs, terminal scoring rule, observable diagnostics, transition type, and same-world branch outcomes. Knowledge cards use closed-set evidence; reasoning cards use checker-verifiable constraints; observation cards use paired visible and occluded fields; tool cards use deterministic stateful interfaces; permission cards change action scope while holding facts fixed; context cards remove or summarize evidence while preserving the requested target. The task carriers follow the semantics of MMLU, MATH, MMMU, and ToolSandbox (Hendrycks et al., 2021a;b; Yue et al., 2024; Lu et al., 2025).

The policy-facing record excludes the response-mode identity, unchosen outcomes, future diagnostics, and evaluator fields. A shared random draw is used across branches of one card so that intervention differences are paired. Parse failures count as unsuccessful terminal outcomes.

F.2 Configurations and Transitions

The main configurations are GPT-5.2, Llama 4 Maverick, GPT-5.6 Sol, Gemini 3.1 Pro, and DeepSeek V4 Pro, with release information drawn from the respective model documentation. These configurations provide distinct competence, diagnostic, tool, and adaptation characteristics; comparisons concern the decision study rather than a general capability ranking.



Figure 5: Sensitivity of normalized utility. Bands are paired 95% card-bootstrap intervals; amber markers denote the validation-selected setting. Moderate choices are stable, while extreme compression, forced matching, diagnostic dominance, or cost weighting degrade the policy.

Stable transitions preserve response incidence and action efficacy. Incidence-only transitions move b; efficacy-only transitions move D; joint transitions move both. Tool-fault transitions change tool reliability with task facts fixed. Access/context transitions change permissions, context availability, or legal branches. Structural tests additionally permute, merge, split, or introduce a response mode and corrupt diagnostics.

F.3 Metrics and Uncertainty

Normalized utility is terminal success plus non-success utility minus branch cost, scaled to a common range. Oracle regret is the same-card gap to the best legal branch. Action-value MAE measures all supported branch predictions. Held-out gain MAE evaluates predicted intervention improvement after excluding a model, drift family, or task–action composition during fitting. Alignment accuracy is computed only for persistent states; unknown-state rate measures rejection on emerged states. Intervals use paired resampling over cards, preserving all methods and branches of a card in the same draw.

F.4 Sensitivity and Uncertainty

Figure 5 varies the four choices most closely tied to the scientific mechanism. Too few modes collapse distinct rescue directions; too many reintroduce variance and weakly separated signatures. A low rejection threshold forces uncertain transport, whereas an overly high threshold discards reusable states. Diagnostics help until their auxiliary loss begins to dominate the action objective. The cost sweep changes both intervention frequency and utility, with the intermediate policy preferred under the study's utility scale.

Two additional controls vary factor regularization and minimum action coverage. Utilities for regularization $\{0, 10^{-3}, 10^{-2}, 10^{-1}\}$ are $\{.721, .727, .730, .724\}$; utilities for coverage floors $\{.02, .05, .10, .20\}$ are $\{.713, .724, .730, .728\}$. The low-coverage loss is the empirical counterpart of the $p_{\min}$ dependence in Proposition 1 and the non-identifiability example above.

Table 6: Normalized utility by configuration. Bold marks the strongest deployable method in each row; Oracle is an evaluator ceiling.

Configuration	Scalar	StationaryVoI	Direct-AV	LowRank-AV	SCOPE	Oracle
GPT-5.2	.666	.686	.693	.696	**.698**	.765
Llama 4 Maverick	.643	.663	.681	.687	**.692**	.742
GPT-5.6 Sol	.728	.747	.753	.759	**.765**	.824
Gemini 3.1 Pro	.711	.733	.748	.756	**.763**	.806
DeepSeek V4 Pro	.691	.712	.721	.728	**.734**	.786
Mean	.688	.708	.719	.725	**.730**	.785

Table 7: Normalized utility by transition. Direct-AV wins in the stable condition; the largest SCOPE gain occurs when incidence and efficacy move together.

Transition	History	Tool health	Belief only	Response only	Direct-AV	LowRank-AV	SCOPE
Stable	.752	.748	.750	.749	**.755**	.753	.747
Incidence only	.718	.704	.736	.706	.728	.734	**.739**
Efficacy only	.707	.719	.699	.731	.730	.732	**.736**
Joint drift	.704	.708	.690	.701	.717	.726	**.740**
Pure tool fault	.707	.734	.704	.732	.733	.733	**.736**
Access/context	.711	.704	.714	.716	.719	.724	**.733**

G Complete Results

G.1 Across Model Configurations

Table 6 shows that action conditioning is the larger improvement: Direct-AV consistently improves over StationaryVoI. LowRank-AV captures part of the remaining low-data gain; SCOPE adds another .002–.007 depending on the configuration. This variation argues against treating operational response modes as a universal replacement for direct value prediction.

G.2 Transition Structure

The one-component ablations match the decomposition without mirroring it perfectly. Belief-only is strongest among ablations for incidence drift, and response-only is strongest for efficacy drift. Under joint drift, their utilities fall to .690 and .701, while SCOPE reaches .740. Pure tool faults are different: the directly observed health signal makes the specialist baseline nearly sufficient.

G.3 Data Efficiency

Table 8: Post-transition feedback study. Direct-AV catches SCOPE in utility once the unrestricted heads receive sufficient feedback.

Feedback Episodes	Utility			Value MAE		
	Direct-AV	LowRank-AV	SCOPE	Direct-AV	LowRank-AV	SCOPE
25	.660	.669	**.674**	.194	.178	**.168**
50	.676	.686	**.692**	.176	.159	**.148**
100	.695	.704	**.711**	.154	.139	**.127**
200	.712	.718	**.724**	.132	.119	**.108**
400	.727	.731	**.733**	.111	.102	**.095**
800	**.742**	.739	.740	.089	.088	**.086**

The reversal at 800 episodes is important. SCOPE still has slightly lower value error, but its small residual factorization bias changes a few close decisions. Direct-AV converts its additional flexibility into higher utility. This separates prediction fit from decision quality near cost-sensitive thresholds.

G.4 Alignment, Merge, Split, and Emergence

Table 9: State-structure results. Alignment accuracy is measured on persistent states; unknown rate is the rejection rate for the designated unmatched component.

Structure	Oracle	SCOPE	LowRank	Direct	No align	Align. acc.	Unknown rate
Permutation	.765	**.751**	.726	.738	.704	.93	.04
Merge	.758	**.742**	.730	.736	.709	.84	.12
Split	.756	**.737**	.728	.734	.701	.79	.18
New state	.749	.716	.719	**.728**	.695	.76	.78
Diagnostic noise	.752	.726	.724	**.731**	.706	.68	.14

Permutation preserves the assumptions of Proposition 2 and produces the cleanest gain. Merge and split reduce separation, narrowing the advantage. A new state creates approximation error: rejecting it is safer than forcing a semantic match, but Direct-AV does not inherit an incorrect old column and therefore wins. Diagnostic noise similarly erodes the signature margin.

H Failure Boundaries

Table 10: Pre-specified boundaries. The table reports normalized utility, not a claim that SCOPE should dominate every condition.

Condition	Best comparator	Comparator	SCOPE	Interpretation
Stable	History	.752	.747	Adaptation overhead is unnecessary
Parallel rescue	StationaryVoI	.724	.722	Action contrasts contain little cause information
High intervention cost	Scalar	.703	.699	Rescue cannot recover its cost
Environment dominated	History	.681	.678	Transient noise is not persistent capability drift
Low overlap	History	.694	.688	Changed response cells remain unidentified
Noisy diagnostics	StationaryVoI	.701	.699	State attribution adds little information
Pure tool fault	Tool health	.734	.736	A specialist observable nearly suffices
Joint drift	Tool health	.708	.740	Both incidence and efficacy must adapt

The first six rows show different reasons not to use the factorization. Stable and environment-dominated worlds do not require structural adaptation. Parallel rescue and high cost collapse the decision distinctions established by Corollary 1. Low overlap violates identification, while noisy diagnostics damage alignment. Pure tool fault is the positive boundary: SCOPE remains competitive, but a much simpler observable rule already captures the useful change.

I Case Analyses

I.1 Knowledge versus Permission at the Same Confidence

Two cards have answer success probability .62. One lacks a fact that retrieval can supply; the other contains the fact but denies execution permission. Scalar routing treats them identically and obtains the correct rescue on .50 of matched pairs. Direct-AV reaches .70 after observing action-specific outcomes. SCOPE reaches .81 by assigning soft mass to two modes whose rescue signatures exchange the retrieval and escalation columns. The .09 gap to the oracle reflects imperfect mode inference rather than confidence error.

I.2 Efficacy Drift without Incidence Drift

Before a tool upgrade, verification rescues .74 of the relevant failures. After the upgrade it rescues .39, while answer success remains .71. A history model and a belief-only update see little reason to change. Signature comparison reveals movement in the verification column, and SCOPE shifts toward replanning. Direct-AV eventually makes the same switch, but its independent verification head needs more selected post-change outcomes. This case is why the paper distinguishes efficacy from incidence rather than treating every accuracy change as capability drift.

I.3 A New State Breaks the Old Geometry

Closing visual capability while changing the output requirement produces a response signature not represented before the transition. Nearest semantic matching maps it to an evidence-gap mode and overuses retrieval. SCOPE rejects most such matches and uses its unknown-state fallback, but detection is imperfect. Direct-AV obtains utility .728 versus .716 for SCOPE. The failure is not repaired by renaming the state: the old factorization lacks the new action-response direction.

J Additional Reporting Details

All methods use the same legal actions, task information, diagnostics, selected outcome feedback, and cost accounting. The oracle alone accesses all branch outcomes and is reported only as a ceiling. Hyperparameters and rejection thresholds are selected on validation transitions and then fixed for held-out model, drift, and task–action tests. Complete item rows retain card identifiers so that paired intervals, utilities, and aggregate tables can be recomputed without treating multiple branches of the same card as independent observations.

The study concerns one-step selection among fixed terminal recovery policies. A multi-step planner could expose each policy as an option, but learning or revising the internal action sequence would require a different state and regret analysis. Likewise, semantic explanations for response modes are descriptive aids; the operational object is their action–diagnostic signature.