TRACEPACT: CONTRACT-INDEXED REPLAY EFFECTS WITH SUPPORT-CONDITIONED EMISSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Local replay scores drive failure localization and corrective updates, but their meaning depends on which replacement is sampled, how much of the trace is rerun, and which terminal event is measured; weak two-arm support can make an equally precise-looking scalar unusable. TRACEPACT makes these choices part of the output: an estimate-or-refuse record binds a typed catalog-average effect to its replay contract, five measured support diagnostics, and calibrated direction-agreement confidence. The record is instantiated by randomized keep-versus-replace logging, a replacement-aware cross-fitted doubly robust learner, and a factorized emission predicate. With identical 5,000-replay budgets on ToolBench and API-Bank, TRACEPACT raises task success from 70.1% to 71.4% and emitted top-1 localization from 53.8% to 62.8%, while reducing direction-agreement ECE from 0.084 to 0.038 relative to forced typed replay. At matched 91.6% coverage, it reaches 57.6% all-request localization versus 54.4% for calibrated empirical-score filtering, a paired 3.2-point gain (95% CI [0.8, 5.6]); on a disjoint 300-request study, all-request top-1 accuracy is 54.7% versus 49.0% with catalog-matched replay labels and 54.3% versus 48.0% with blinded labels. Controlled catalog variants change the fitted top-ranked segment in 27.3% of 13,842 comparisons and its effect sign in 16.8%, while backend support erosion raises abstention from 8.4% to 29.3% at fixed assignment propensity. These results establish a concrete interface for comparing, selecting, and consuming local replay effects in the evaluated agent-trace regimes.

1 INTRODUCTION

Replay-based attribution is increasingly consumed as an action: the highest-scoring segment is inspected, replaced, or used to weight an offline update. Yet a failed tool-use trace has no catalog-free "culprit." Replacing a retrieval with an item from the same index, suppressing it, and resampling an action define different interventions; regenerating three descendants and rerunning an entire suffix define different causal boundaries. A bare scalar hides these choices and also hides whether both randomized arms were observed often enough near the queried trace.

TRACEPACT changes the reporting unit from a score to a self-identifying effect record. Its contract $\chi = (k, \mathcal{P}_k, W, R)$ binds the segment type, replacement distribution, replay window, and terminal checker. The record carries the catalog-average local effect and repair lift, the diagnostics that govern emission, calibrated direction-agreement confidence, and either an estimate or an explicit insufficient-support status. Records can be pooled as effect estimates only when contracts align; selective-risk comparisons also align support policies. Ranking prespecified type-specific catalogs within an episode is instead a decision comparison on their common terminal-success scale.

The statistical construction follows the semantics of that record. Randomized keep-versus-replace logging records the sampled replacement identity; nuisance heads condition on that identity before marginalizing over the catalog. Cross-fitted selection-adjusted pseudo-outcomes feed a held-out conditional-effect learner, while a separate validation split calibrates direction agreement and selects the uncertainty threshold. Section 3.2 distinguishes population identification support from the finite-sample effective-count and uncertainty diagnostics, and Section 4.2 states the identification result for the contract-indexed effect. The resulting interface is the technical contribution: intervention identity, numerical estimation, and evidence for emission remain coupled from replay through use.

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

The evidence tests each link in this interface. Under the same 5,000 replays, catalog, checker, folds, and update, TRACEPACT improves emitted top-1 localization by 9.0 points and task success by 1.3 points over forced typed replay, while direction-agreement ECE falls by 0.046. At identical 91.6% coverage, its end-to-end all-request localization exceeds calibrated filtering of empirical typed-replay scores by 3.2 points (95% CI [0.8,5.6]). Known-effect synthetic traces separate effect recovery from selection; cumulative components locate the conditional-estimation and emission gains; catalog and backend shifts probe contract identity and support response. The ordering then persists on disjoint catalog-matched and blinded endpoints, WebArena traces, a smaller acting policy, and an AWAC consumer. Together these analyses support contract-indexed, estimate-or-refuse replay as a reliable input to agent diagnosis and updating.

2 RELATED WORK AND POSITIONING

Recent agent-trace diagnosis makes step-level failure localization an explicit computational problem. AgentRx derives guarded constraints and asks a judge to identify the first critical violation (Barke et al., 2026); AgenTracer learns agent- and step-level attribution from counterfactual fault-injection traces (Zhang et al., 2026); and REFLECT tests a diagnosis-specific patch and feeds a successful outcome flip back into localization (Lin et al., 2026). Causal Agent Replay (CAR) estimates effects under stochastic forward reruns and adds a point-of-commitment rule and interaction-aware Shapley credit (Shah, 2026); CausalFlow likewise couples counterfactual attribution with candidate repair (Bonagiri et al., 2026). For cooperative multi-agent training, C3 freezes transcript context and computes message-level leave-one-out advantages under fixed-continuation replay (Chen et al., 2026). These methods return constraint evidence, learned attribution, verified repairs, stochastic rerun effects, interactions, or training credit. TRACEPACT instead makes the returned statistical object self-identifying: the output is a catalog-marginal single-segment effect whose intervention fields, empirical support, direction confidence, and refusal status travel together.

Feature attribution and hindsight credit supply additional comparators. Integrated gradients and SHAP rank inputs under model-relative perturbations (Sundararajan et al., 2017; Lundberg & Lee, 2017; Shapley, 1953), while generalized advantages propagate realized return through a trajectory (Schulman et al., 2016). Importance weighting, contextual replay, doubly robust evaluation, and double machine learning provide the estimation tools for randomized or logged alternatives (Precup et al., 2000; Dudík et al., 2011; Li et al., 2011; Swaminathan & Joachims, 2015; Jiang & Li, 2016; Thomas & Brunskill, 2016; Chernozhukov et al., 2018). TRACEPACT uses these tools inside a query whose catalog, replay boundary, outcome, and emission predicate remain visible in the output.

Reject-option learning and selective prediction study risk as coverage varies (Bartlett & Wegkamp, 2008; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019; Goren et al., 2024); calibration relates reported confidence to observed reliability (Guo et al., 2017; Lakshminarayanan et al., 2017; Minderer et al., 2021). Our coverage-matched selective baselines apply count, ensemble-confidence, or calibrated-confidence filters to the same empirical typed-replay score. The comparison therefore tests the full conditional-estimation and support-aware pipeline against generic filtering of that score; the cumulative ablation in Table 4 separately exposes the estimator, calibration, effective support, uncertainty, feasibility, and observability steps.

ToolBench, API-Bank, and WebArena expose long-horizon tool-use failures in the three evaluated domains (Qin et al., 2024; Li et al., 2023; Zhou et al., 2024); Mind2Web and AgentBench provide related web and interactive-agent settings (Deng et al., 2023; Liu et al., 2024). WebGPT, Gorilla, Reflexion, Voyager, and SWE-agent motivate repair from execution feedback (Nakano et al., 2021; Patil et al., 2024; Shinn et al., 2023; Wang et al., 2024; Yang et al., 2024). Our evaluation tests whether a local replay conclusion remains identifiable, supported, and useful when its catalog or backend changes.

3 CONTRACT-INDEXED REPLAY EFFECTS

3.1 CONTRACT IDENTITY AND REPAIR LIFT

An agent trace $\tau = (V, E)$ is a directed acyclic graph with decision v^d, retrieval v^r, tool-output v^t, and observation v^o nodes. Edges encode local dependencies and the terminal node carries benchmark

Table 1: Interface-level comparison with the closest agent-trace analysis objects. The columns expose differences in intervention identity, validity handling, and the object returned to a consumer.

Method family	Intervention or evidence	Effect target / validity handling	Output
AgentRx / AgenTracer	constraints or injected trace faults	critical step without a catalog-marginal effect	step / agent
REFLECT / CausalFlow	selected repair and rerun	repair-specific outcome change; no support refusal	repaired locus
CAR	stochastic forward intervention	single-step or interaction effect with intervals	locus / credit
C3	fixed-context message replacement	message advantage under fixed continuation	training credit
Selective typed replay	one typed catalog draw	empirical score plus scalar threshold	score / refusal
TRACEPACT	typed catalog draw; bounded local rerun	catalog-average effect plus five support tests	effect record / refusal

return $R \in \{0, 1\}$. For node v of type k, $D_W(v)$ contains the next $W = 3$ descendants reopened by replay. A query

$$\chi_v = (k, \mathcal{P}_k, W, R), \qquad q_v = (x_v, \chi_v)$$

binds typed context x_v to a local-effect contract χ_v: replacement catalog $\mathcal{P}_k$, replay window W, and terminal checker R. Let A_v denote the replacement identity. For $z = 0$, $A_v \sim \mathcal{P}_k(\cdot \mid x_v)$; for $z = 1$, its distribution is a point mass on a keep symbol. The full-data potential outcome $Y_{\chi_v}(\tau, v, z, a) \in \{0, 1\}$ is terminal success under the contract's transition-consistent continuation for arm z and identity a. It is defined independently of whether the replay logger observes that endpoint. Writing $Y_{\chi_v}(\tau, v, z) = Y_{\chi_v}(\tau, v, z, A_v)$, with the expectation integrating over the arm-specific identity distribution, the contract-indexed local effect is

$$\delta_{\chi_v}(x_v) = \mathbb{E}[Y_{\chi_v}(\tau, v, 1) - Y_{\chi_v}(\tau, v, 0) \mid x_v]. \tag{1}$$

The outcome remains end-to-end benchmark success; W bounds controlled re-execution, not the reward horizon. For failure repair, $\rho_{\chi_v}(x_v) = -\delta_{\chi_v}(x_v)$ is the expected success lift from replacing the original payload. Localization ranks $\widehat{\rho}_{\chi_v}$, while downstream weighting uses $|\widehat{\delta}_{\chi_v}|$.

The catalog fixes the scientific question. $\mathcal{P}_d$ samples policy actions at temperature $T = 1.5$; $\mathcal{P}_r$ retrieves alternatives from the same frozen index and interface; $\mathcal{P}_t$ selects cached responses with the same API signature; and $\mathcal{P}_o$ substitutes null tokens within the same observation channel. Only contract-aligned records estimate the same effect; ranking type-specific candidates within an episode is a downstream decision on their common terminal-success scale.

Equation 1 is the record's statistical target: the replacement-arm potential outcome is marginalized over one draw from $\mathcal{P}_k(\cdot \mid x_v)$. We evaluate that target directly through RMSE and rank correlation on generator-known effects. Real-domain top-1 localization is a downstream decision endpoint: it asks whether ranking $\widehat{\rho}_{\chi_v}$ selects the segment identified by a prespecified repair rubric or a catalog-matched fixed replay bank. Section 5 defines and Section 6 evaluates these endpoints separately so that effect recovery and repair localization do not substitute for one another.

3.2 ACCEPTED SUPPORT REGION

Binary arm assignment is randomized: $Z_v = \mathbf{1}[v \text{ remains original}]$ has logged propensity $e_\chi(x_v) = 1 - p_{\text{perturb}} = 0.88$, with $p_{\text{perturb}} = 0.12$. Catalog draws in the replacement arm are randomized according to the recorded $\mathcal{P}_k$. Write $Y_v(z, a) \equiv Y_{\chi_v}(\tau, v, z, a)$, and for the logged replay write $C_v \equiv C_v(Z_v, A_v)$ and $Y_v \equiv Y_v(Z_v, A_v)$. The indicator $C_v(z, a)$ records whether execution and logging expose $Y_v(z, a)$ through a transition-consistent state. Thus C is a post-assignment outcome-observation indicator, not part of the definition of Y. Identification uses conditional replay observability,

$$C_v(z, a) \perp Y_v(z, a) \mid x_v, \chi_v, a, \qquad q_{\chi,z}(x, a) = \Pr\{C_v(z, a) = 1 \mid x_v = x, \chi_v = \chi, a\}.$$

For support reporting, $\bar{q}_{\chi,z}(x) = \mathbb{E}_{A\sim\mathcal{P}_{\chi,z}}\{q_{\chi,z}(x,A)\}$ averages this probability over the arm-specific identity distribution. Conditioning before catalog marginalization prevents a well-observed replacement from masking a systematically unobserved identity. Conditional replay observability remains an identifying assumption; hidden-state omission probes robustness to incomplete conditioning rather than verifying the independence.

A support policy $\gamma = (\epsilon_q, n_{\min}, \kappa)$ sets the replay-observability floor $\epsilon_q = 0.08$, per-arm effective-count floor $n_{\min} = 3$, and validation-selected uncertainty threshold κ. All diagnostics are computed without audit labels. The contextual neighborhood $\mathcal{N}_\chi(x)$ is the 64 nearest training-fold requests that share the segment type and contract family, measured by cosine distance in the frozen trace-embedding space. Within arm z, observability-adjusted weights retain the recorded identity A_i:

$$w_{iz}(x) = \frac{\mathbf{1}\{i \in \mathcal{N}_\chi(x), Z_i = z, C_i = 1\}}{\widehat{q}_{\chi,z}(x_i, A_i)}, \qquad n_{\text{eff},z}(x,\chi) = \frac{\{\sum_i w_{iz}(x)\}^2}{\sum_i w_{iz}(x)^2}.$$

The neighborhood size and metric are fixed before outcome analysis. For the effect ensemble, uncertainty is

$$\widehat{\sigma}^2_\chi(x) = \mathrm{Var}_{f,s}\{\widehat{\delta}^{(f,s)}_\chi(x)\} + \mathrm{median}_{f,s}\{\widehat{V}^{(f,s)}_{\text{epboot}}(x)\},$$

combining fold/seed dispersion with episode-bootstrap sampling variance. Validation folds choose κ on the risk–coverage grid by maximizing correct-direction mass, $\text{coverage} \times (1 - \text{direction error})$, with lower direction-agreement ECE as the tie-breaker. TRACEPACT then records five separate certificate quantities:

1. catalog feasibility F_v, verifying that a type-correct replacement exists;
2. realized terminal observability $C_v = C_v(Z_v, A_v)$ for the logged replay and identity-aware heads $q_{\chi,z}(x,a)$, requiring catalog-marginal $\bar{q}_{\chi,z}(x) \geq \epsilon_q$ for both arms;
3. per-arm contextual effective counts $n_{\text{eff},z}(x_v,\chi)$ in the effect learner's neighborhood;
4. logged arm propensity $e_\chi(x_v)$, retained as a randomized-design positivity diagnostic; and
5. conditional-effect uncertainty $\widehat{\sigma}_\chi(x_v)$ and held-out calibration error.

The population identification support and finite-sample statistical gate are

$$\begin{aligned}
\mathcal{S}_\chi &= \{x : e_\chi(x) \in (0,1),\ \min_z \bar{q}_{\chi,z}(x) > 0\},\\
E_v &= F_v C_v \prod_{z\in\{0,1\}} \mathbf{1}\{\widehat{q}_{\chi,z}(x) \geq \epsilon_q,\\
&\qquad n_{\text{eff},z}(x,\chi) \geq n_{\min}\}\\
&\qquad \mathbf{1}\{e_\chi(x) \in (0,1),\ \widehat{\sigma}_\chi(x) \leq \kappa\}.
\end{aligned}$$

Here $\widehat{q}_{\chi,z}(x)$ in the gate denotes the catalog-marginal $\widehat{\bar{q}}_{\chi,z}(x)$, and every gate quantity is evaluated at $x = x_v$. The first line is deterministic population support; the second directly defines the finite-sample query-level emission indicator and includes realized F_v and C_v. If $C_v = 0$, the full-data outcome remains latent, no numerical value is imputed, and the query is refused. Every failed factor and its measured value remains in the record. The final cumulative-ablation step is explicit: the uncertainty-only row lacks query feasibility and realized observability checks.

Appendix Figure 3 reports a context-support index: the mean smaller-arm effective support after replay-observability weighting, normalized to a validation reference and clipped to $[0, 100]$. The index falls from 92% to 42% under backend shift; coverage is separately determined by query checks, hard floors, and the uncertainty gate, while randomized propensity remains 0.88.

4 TRACEPACT: CERTIFICATION AND EMISSION

4.1 REPLAY CONTRACT

The logger samples one node, records x_v, type, arm probability, replacement identity, environment snapshot, and terminal outcome. Ancestors remain fixed. At v, replay either preserves the payload or inserts exactly one catalog draw. The acting policy then regenerates the first three descendants against

the same frozen retrieval index, signature-matched tool cache, and benchmark sandbox or mirror. At the $W = 3$ boundary, the regenerated state is serialized into the benchmark state representation; intervention control is released and the original acting policy and environment transition normally from that state until the benchmark's terminal checker fires. The old suffix is never pasted onto an incompatible state. Thus the intervention surface is local while downstream consequences reach the terminal outcome.

4.2 Cross-fitted conditional DR learner

For each contract family χ, five-fold cross-fitting estimates identity-aware outcome heads $\widehat{\mu}_{\chi,z}(x,a) = \mathbb{E}[Y \mid x, Z = z, C = 1, \chi, A = a]$ and observability heads $\widehat{q}_{\chi,z}(x,a)$. For the keep arm, a is the fixed keep symbol. Before contrasting arms, the heads are marginalized over the corresponding recorded catalog, for example $\widehat{\bar{\mu}}_{\chi,0}(x) = \mathbb{E}_{A\sim\mathcal{P}_k(\cdot|x)}\widehat{\mu}_{\chi,0}(x,A)$. The logged randomization probability $e_\chi(x)$ is used directly. For an observation held out from all nuisance heads, the replacement-aware selection-adjusted DR pseudo-outcome is

$$\widehat{\psi}_v = \widehat{\bar{\mu}}_{\chi,1}(x_v) - \widehat{\bar{\mu}}_{\chi,0}(x_v) + \frac{C_v Z_v\{Y_v - \widehat{\mu}_{\chi,1}(x_v,A_v)\}}{\widehat{q}_{\chi,1}(x_v,A_v)e_\chi(x_v)} - \frac{C_v(1-Z_v)\{Y_v - \widehat{\mu}_{\chi,0}(x_v,A_v)\}}{\widehat{q}_{\chi,0}(x_v,A_v)\{1-e_\chi(x_v)\}}. \quad (2)$$

Equation 2 is an orthogonal pseudo-outcome, not yet a per-query estimate. A held-out contract-family effect head maps it to

$$\widehat{\delta}_\chi(x) = \widehat{g}_\chi(x) \approx \widehat{\mathbb{E}}[\widehat{\psi}_v \mid x_v = x, \chi_v = \chi], \quad (3)$$

using the same four-layer architecture as the nuisance heads. Each correction term is an inverse-observability-weighted residual for the logged replacement identity; under conditional replay observability it identifies the corresponding identity-specific arm mean, and catalog marginalization yields $\delta_\chi(x)$ (Appendix A). Nested folds ensure that no query trains its own nuisance, effect, or calibration head. An ensemble across folds and seeds supplies epistemic variation, and an episode bootstrap supplies sampling variation. On validation folds, the reference contrast is $\widetilde{\delta}_\chi(x) = \sum_i w_{i1}Y_i/\sum_i w_{i1} - \sum_i w_{i0}Y_i/\sum_i w_{i0}$ in the same 64-neighbor set; it uses the same catalog, W, and R, and at most one logged replay per source episode. The raw direction statistic is the fraction of ensemble members agreeing with the modal contract-effect sign. Isotonic calibration maps it to the indicator $\mathbf{1}\{\mathrm{sign}(\widehat{\delta}_{\chi,\mathrm{modal}}) = \mathrm{sign}(\widetilde{\delta}_\chi)\}$; equivalently, both signs may be negated for repair lift. This target is independent of localization annotations and is not a continuous-effect calibration target.

Proposition 1 (Contract-effect identification and consistency) *Suppose replay consistency, randomized arm and catalog assignment, conditional replay observability as stated above, and bounded outcomes. On any compact $K \subset \mathcal{S}_\chi$, assume e_χ, $1 - e_\chi$, and $q_{\chi,z}(x,a)$ are uniformly bounded away from zero. If, for each arm and identity, either $\mu_{\chi,z}(x,a)$ or $q_{\chi,z}(x,a)$ is consistently estimated, catalog marginalization gives $\mathbb{E}\{\psi(O;\eta_\chi) \mid x,\chi\} = \delta_\chi(x)$. If the conditional cross-fitted pseudo-outcome bias is uniformly $o_p(1)$ on K and the effect learner is consistent for $\mathbb{E}(\psi \mid x,\chi)$, then $\widehat{\delta}_\chi$ is uniformly consistent for Eq. 1 on K.*

The proposition governs the numerical effect, while E_v is the finite-sample reporting policy. Logged propensities document assignment, and $\widehat{q}$ with effective counts diagnoses observed positivity. Replay consistency and conditional observability remain assumptions; frozen transition semantics and hidden-state omission respectively support the intervention definition and probe robustness to incomplete conditioning. Appendix A gives the identification argument.

4.3 Emission and downstream use

The requested query is q_v; it emits when $E_v = 1$. Its record is

$$r_v = (q_v, \gamma, s_v, \widehat{\delta}_\chi(x_v), \widehat{\rho}_\chi(x_v), c_v, \mathrm{status}_v),$$

Here s_v contains certificate diagnostics, c_v is calibrated confidence, and status is "estimate" or "insufficient support." A refused record sets the numerical effect and repair-lift fields to null. For an audited episode, all candidate node types use their own $\mathcal{P}_k$ but share W and R, so their repair lifts are changes on the same terminal-success scale. The request emits if at least one candidate

emits and selects $\arg\max_{v:E_v=1} \widehat{\rho}_{\chi_v}(x_v)$; it refuses only when every candidate refuses. Coverage is emitted requests divided by all requests; all-request localization utility is correct emissions divided by all requests; full abstention contributes zero. A top-1 prediction is correct when it matches the prespecified endpoint for that evaluation.

Direction agreement, localization, and downstream weighting are distinct consumers of the same record. Raw confidence is the ensemble fraction defined above; because $\rho = -\delta$ negates every member, the same fraction applies to repair direction. Method-specific isotonic calibration is fit only on the common validation folds against the out-of-fold replay contrast. Direction-agreement ECE pools emitted records across contract families after calibration and averages the absolute confidence–agreement gap over ten equal-width bins; exact zero contrasts use $\mathrm{sign}(0) = 0$. It is a replay-agreement proxy, not an effect-magnitude or localization metric. Localization uses the selected segment, while the updater uses $|\widehat{\delta}|$.

One downstream consumer performs a single weighted behavior-cloning update. Decision target $a_u^\star$ is the successful sampled replacement when perturbing u flips failure to success and otherwise the logged action. Refused records contribute no numerical weight. Supported non-decision effects are routed uniformly to descendant decisions in $D_W(v)$:

$$\mathcal{L}_{\text{weighted}} = \sum_{u \in V^d} \left(|\widehat{\delta}_{\chi_u}(x_u)| + \sum_{v \in \mathcal{A}_W(u)} \alpha_{v \to u} |\widehat{\delta}_{\chi_v}(x_v)| \right) \{ -\log \pi_\theta(a_u^\star \mid x_u) \}. \tag{4}$$

Here $\mathcal{A}_W(u) = \{v : u \in D_W(v)\}$ and $\alpha_{v \to u}$ is uniform over supported descendant decisions. All matched-cost methods share targets, routing, replay traces, and update steps, isolating the emitted record used as weight.

5 Experimental Protocol

We evaluate 2,800 ToolBench (Qin et al., 2024), 1,400 API-Bank (Li et al., 2023), and 1,850 WebArena-Trace (Zhou et al., 2024) episodes. WebArena-Trace denotes benchmark episodes represented in the same typed event-graph format and replayed in benchmark mirrors; official task success remains the terminal checker. The acting policy is Llama-3.1-70B; four-layer, width-256 outcome, observability, and effect heads consume frozen Llama-3.1-8B trace embeddings (Grattafiori et al., 2024). Real-domain logging yields 15,000 replay episodes. Every matched comparison reuses one common 5,000-replay subset per seed, the same episode-level folds, and one update; the remaining observations enter only analyses explicitly labeled full-pool. Runs use eight NVIDIA A100-80GB GPUs, CUDA 12.4, PyTorch 2.3.1, Adam at 2×10^{-4}, batch size 128, and 80K steps. Total compute is 420 A100-GPU-hours.

The baseline suite contains uniform weighting, integrated-gradient saliency, 200-permutation Shapley approximation, hindsight advantage, leave-one-out, matched-cost leave-k-out with $k = 3$, off-policy advantage, and forced typed replay. Forced typed replay uses the same catalog and empirical mean terminal-success deltas and emits every feasible query. Validation folds fit each selective statistic and calibrator. For coverage-matched analysis, a label-blind rank cutoff on held-out scores makes each rule emit the same number of evaluation requests as TracePact; no evaluation labels enter this cutoff. Table 3 supplies the end-to-end comparison against generic empirical-score selection, and Table 4 separates the empirical-to-conditional-DR improvement from the fixed-order emission path. No method receives extra replay outcomes, audit labels, catalog edits, or update steps.

The split unit is the source episode. In each outer fold, replay outcomes inherit their source-episode identity; nuisance and effect heads use only training episodes, isotonic calibration and κ selection use a disjoint validation partition within that training side, and the weighted update uses supported training-fold records. Post-update task success is scored once on the untouched outer fold. Audit labels never enter fitting, calibration, gate selection, or updating. Five seeds reuse split identities and vary logger/model randomness. Confidence intervals resample source episodes jointly across methods within each seed, recompute the seed-level paired contrast, and then aggregate the five seed contrasts; they are conditional on the five fitted seeds and quantify episode-sampling rather than across-training-run uncertainty. Appendix Table 5 consolidates pool ownership.

The primary localization set contains 417 prespecified requests from the ToolBench+API-Bank slice of the 612-episode pool, one per failed episode. Two benchmark-familiar author annotators independently selected the earliest typed segment whose catalog-admissible replacement could avert terminal failure; they saw neither method scores nor support status, used earliest-segment tie breaking, and adjudicated only after recording independent labels. Cohen's $\kappa = 0.79$ is measured before adjudication. The independent 300-request set—120 ToolBench, 80 API-Bank, and 100 WebArena—has two endpoints. Its catalog-matched replay endpoint pre-samples three admissible replacements per segment, averages their observed terminal-success lift relative to keep, and selects the segment with greatest mean lift. Its blinded endpoint uses majority vote from three non-author, benchmark-familiar annotators applying the minimal-success-repair rubric with the same tie rule. The replay endpoint is a three-draw Monte Carlo decision target aligned with the catalog-average lift; the annotation endpoints measure the complementary earliest-actionable-repair decision rather than Eq. 1 itself.

Annotation concerns benchmark traces and task outcomes, collects no annotator behavior or personal data, and supports only the evaluation labels above. Reporting the primary adjudication, catalog-matched replay endpoint, and disjoint non-author endpoint separately exposes agreement across three evidence sources; Fleiss' κ is computed before majority resolution. All interventions run in benchmark sandboxes or mirrors, never against live users or services.

6 RESULTS

6.1 MATCHED-COST AND MATCHED-COVERAGE COMPARISONS

Table 2 fixes the intervention family and downstream consumer. Relative to forced typed replay, TRACEPACT adds 1.3 task-success points and 9.0 emitted top-1 localization points while reducing direction-agreement ECE by 0.046. The complete outcome vector favors the contract-indexed record under the same update, and the localization difference is substantially larger than the downstream aggregate change; Table 9 supplies the broader attribution context.

At identical coverage, the full pipeline produces 240 correct records versus 227 for calibrated filtering of empirical replay scores. The paired five-seed advantage is 3.2 all-request localization points (95% CI [0.8, 5.6]); row means also differ by 0.6 task-success points and 0.009 direction-agreement ECE. This is the primary end-to-end selective comparison, while Table 4 identifies how much improvement is already present before refusal. Pooled integer fractions and paired seed-level estimates use different aggregations, explaining tenth-point rounding differences. Episode-bootstrap 95% CIs are 91.6% [89.0,94.0] for TRACEPACT coverage and 0.038 [0.033,0.043] for its direction-agreement ECE, compared with 0.047 [0.041,0.053], 0.053 [0.047,0.060], and 0.061 [0.055,0.068] for calibrated, raw-confidence, and count-floor filtering.

6.2 CONTRACT IDENTITY UNDER CATALOG SHIFT

We vary $\mathcal{P}_k$ while holding W and R fixed: the conservative and expanded contracts use a subset and a superset of the default type-valid alternatives, respectively, while the held-out shifted contract replaces them with a disjoint type-valid set.

Across 13,842 feasible query–catalog comparisons with numerical effects under both catalog settings, the fitted top-ranked segment changes for 27.3% and the estimated effect sign changes for 16.8%.

Table 2: Matched-cost comparison on ToolBench+API-Bank. All rows preserve the same 5,000 replay episodes, typed catalog, terminal checker, and downstream update. Entries are five-seed means with episode-bootstrap 95% intervals; direction-agreement ECE is emission-only.

Method	Replays	Task succ. ↑	Emit. loc. ↑	Dir.-ECE ↓
Leave-k-out ($k = 3$)	5,000	69.8 ± 0.7	52.1 ± 1.4	0.078
Forced typed replay	5,000	70.1 ± 0.8	53.8 ± 1.6	0.084
Off-policy advantage (GAE)	5,000	69.4 ± 0.8	49.2 ± 1.7	0.091
TRACEPACT	5,000	71.4 ± 0.6	62.8 ± 1.3	0.038

Table 3: Coverage-matched filtering on 417 requests. Count and confidence rows filter empirical typed-replay scores; TRACEPACT uses conditional DR plus the full predicate. "All" is correct emissions per request; every selective row emits 382 records (91.6%). Row intervals are bootstrap 95% intervals; the paired primary contrast is reported below. Direction-agreement ECE is emission-only.

Method	Task	Emit.	All	Cov.	Dir.-ECE	Correct/emitted
Forced typed	70.1 ± 0.8	53.8 ± 1.6	53.8%	100%	0.084	224/417
Count floor	70.5 ± 0.7	57.1 ± 1.5	52.3%	91.6%	0.061	218/382
Raw confidence	70.6 ± 0.7	58.1 ± 1.4	53.2%	91.6%	0.053	222/382
Calibrated selective	70.8 ± 0.7	59.4 ± 1.4	54.4%	91.6%	0.047	227/382
TRACEPACT	71.4 ± 0.6	62.8 ± 1.3	57.6%	91.6%	0.038	240/382

Exact zero uses sign 0, and no magnitude threshold removes near-zero changes; these rates therefore measure sensitivity of the fitted record to its contract. Under the held-out shifted catalog, TRACEPACT retains 6.8 emitted-localization points and a 0.027 direction-agreement ECE advantage over calibrated filtering at matched 73.6% coverage (Appendix Table 7).

6.3 MECHANISM DECOMPOSITION

Table 4: Cumulative matched-cost decomposition on the same requests. Entries are five-seed means with bootstrap 95% intervals; direction-agreement ECE is emission-only. Calibration leaves forced-emission ranking fixed; correct/request is the all-request endpoint.

Configuration	Task	Emit. loc.	Abstain	Dir.-ECE	Correct/request
Typed empirical, forced	70.1 ± 0.8	53.8 ± 1.6	0.0%	0.084	224/417
Conditional DR, forced	70.8 ± 0.7	57.1 ± 1.5	0.0%	0.064	238/417
+ held-out calibration	70.8 ± 0.7	57.1 ± 1.5	0.0%	0.044	238/417
+ effective-support gate	71.0 ± 0.7	59.8 ± 1.4	6.0%	0.041	234/417
+ uncertainty gate	71.2 ± 0.6	61.4 ± 1.4	7.7%	0.040	236/417
+ query checks F_v, C_v (TRACEPACT)	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038	240/417

Conditional DR raises forced-emission localization by 3.3 points and correct/request from 224/417 to 238/417 before any request is refused; calibration then reduces direction-agreement ECE without changing the ranking. The fixed gate sequence trades coverage for a progressively cleaner emitted set, reaching 240/417 correct requests and 62.8% emitted localization. Relative to calibrated forced DR, the complete predicate adds 5.7 [3.0,8.4] emitted-localization points (paired 95% CI). The decomposition assigns complementary roles: conditional DR raises all-request correctness, while the certificate improves emitted-set reliability and ends with two additional correct requests. For readability, the table adds effective support and uncertainty before the query-level F_v, C_v checks, although the definition of E_v lists all factors together.

6.4 HELD-OUT ENDPOINT VALIDATION

TRACEPACT and calibrated selective replay emit 89.7% and 90.7% of the disjoint requests. At these nearby coverages, TRACEPACT's paired emitted-accuracy gains are 7.0 [1.0,12.9] points on replay-confirmed labels and 7.7 [1.5,13.8] on blinded labels; pooled all-request accuracy is 54.7% versus 49.0% and 54.3% versus 48.0%, respectively. Appendix Table 6 reports counts. Pre-adjudication Fleiss' $\kappa = 0.74$ indicates substantial agreement across the blinded annotators.

6.5 STABILITY AND SUPPORT SHIFT

Across $W \in \{2, 3, 5\}$, $\epsilon_q \in \{0.05, 0.08, 0.12\}$, $p_{\text{perturb}} \in \{0.08, 0.12, 0.18\}$, and count floors $\{2, 3, 5\}$, emitted localization spans 60.3–63.3% and direction-agreement ECE spans 0.035–0.047. The selected point combines 62.8% emitted localization, 57.6% all-request utility, 91.6% coverage,

and 0.038 direction-agreement ECE; all-request utility remains within 0.1 points across 91.6–95.0% coverage.

At an $s\%$ backend shift, $s\%$ of replay requests come from held-out backend variants while arm propensity, catalog, W, and R remain fixed. From $s = 0$ to 50, the normalized context-support index falls from 92% to 42%, abstention rises from 8.4% to 29.3%, and direction-agreement ECE moves from 0.038 to 0.093. This trajectory links contextual observability and effective-sample erosion to progressively stricter emission. The zero-shift setting is the ToolBench+API-Bank operating point; WebArena's base abstention is 12.1% (Table 10).

6.6 Effect fidelity, domain difficulty, and transfer

On 5,000 synthetic source traces evaluated with the full 15,000-replay pool, TracePact reduces common-answerable-set effect RMSE from 0.155 ± 0.008 for Shapley approximation to 0.082 ± 0.005 and raises rank correlation from 0.64 ± 0.02 to 0.83 ± 0.02 (Table 8). At the matched 5,000-replay frontier, its RMSE is 0.085 (Figure 1); separating the replay budgets makes the effect-fidelity result directly comparable at both operating points. On WebArena-Trace, TracePact reaches 42.4 ± 0.8 task success, 54.6 ± 1.5 emitted localization, 12.1% abstention, and 0.042 direction-agreement ECE, preserving the diagnostic advantage on longer web interactions (Table 10).

The ordering also transfers across policy scale and updater. With Llama-3.1-8B, TracePact reaches 63.6 ± 0.8 task success and 59.4 ± 1.6 localization, compared with 62.0 ± 0.9 and 50.9 ± 1.7 for forced typed replay; AWAC retains 72.2 ± 0.6 task success and 63.0 ± 1.3 localization (Table 14). Under 30% catalog contamination, the certificate records 9.4% false emissions and 0.066 direction-agreement ECE, versus 22.1% and 0.096 for calibrated filtering (Table 13). These tests connect conditional DR to effect fidelity, the full pipeline to domain and consumer transfer, and the support predicate to controlled refusal under misspecification.

7 Discussion

The held-out catalog shift exposes the operational difference between a detached score and a contract-indexed record: at the same 73.6% coverage, TracePact gains 6.8 emitted-localization points and reduces direction-agreement ECE by 0.027 relative to calibrated filtering. Conditional DR first improves all-request localization; emission factors then concentrate reliable records as support erodes. Agreement across disjoint endpoints, WebArena, policy scale, and updater shows that the separation is not tied to one evidence source. The rule is concrete: pool effect estimates only under aligned contracts and support policies; cross-type localization is a terminal-success decision, not a shared estimand.

The certificate quantifies empirical support for single-segment, observed-state replay with a fixed contract. Replay consistency and conditional observability delimit its interpretation; hidden-state omission probes incomplete conditioning through lower coverage and a narrower RMSE separation, while catalog contamination raises false emissions and direction-agreement ECE. Interacting multi-node repairs define a different contract, and model error remains visible through $\widehat{q}$, dispersion, calibration, and refusal. Across the evaluated regimes, explicit identity, measured support, and controlled degradation support the downstream conclusions without treating the record as a catalog-independent causal label.

8 Conclusion

Agent diagnosis and updating can consume r_v only after its query contract, support policy, diagnostics, confidence, and status are read together. Matched-cost, matched-coverage, known-effect, and disjoint-endpoint evidence shows that this protocol improves localization, calibrates direction agreement, and supplies more effective update weights across the evaluated trace regimes. Catalog and backend shifts show why carrying effect identity and empirical support through the interface is necessary for contract-matched, estimate-or-refuse downstream use.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and revision, presentation organization, formatting checks, and reference normalization during preparation. The reported methods, definitions, and measurements are retained from the existing research record.

REFERENCES

Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI agent failures from execution trajectories. *arXiv preprint arXiv:2602.02475*, 2026.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026.

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.

Xiang Deng et al. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, pp. 1097–1104, 2011.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 297–306, 2011.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.187.

Xiaofeng Lin, Yingxu Wang, Tung Sum Thomas Kwok, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. REFLECT: Intervention-supported error attribution for silent failures in LLM agent traces. *arXiv preprint arXiv:2606.09071*, 2026.

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*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Reiichiro Nakano et al. WebGPT: Browser-assisted question answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pp. 759–766, 2000.

Yujia Qin et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games*, volume 2, pp. 307–317, 1953.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pp. 814–823, 2015.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the LLM agentic systems? In *International Conference on Learning Representations*, 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

A CONTRACT-EFFECT IDENTIFICATION

Let $m_{\chi,z}(x,a) = \mathbb{E}\{Y_v(z,a) \mid x, \chi, a\}$, and let $\mu_z^\star(x,a), q_z^\star(x,a)$ denote probability limits of the fitted heads. Random arm and catalog assignment plus replay consistency connect each observed tuple (z,a) to $Y_v(z,a)$; conditional replay observability gives $\mathbb{E}\{Y_v(z,a) \mid C_v(z,a) = 1, x, \chi, a\} = m_{\chi,z}(x,a)$. The conditional mean of one identity-specific augmented term is

$$\mu_z^\star(x,a) + \frac{q_{\chi,z}(x,a)}{q_z^\star(x,a)}\{m_{\chi,z}(x,a) - \mu_z^\star(x,a)\}.$$

It equals $m_{\chi,z}(x,a)$ if either $\mu_z^\star = m_{\chi,z}$ or $q_z^\star = q_{\chi,z}$. Integrating this identity over the arm-specific catalog and applying it to both randomized arms in Eq. 2 yields $\mathbb{E}\{\psi(O;\eta_\chi) \mid x, \chi\} = \delta_\chi(x)$. Cross-fitting prevents a request from fitting its own nuisance predictions, and positivity bounds control inverse-weight amplification. Under the proposition's uniform conditional-bias and effect-regression conditions, the triangle inequality gives $\sup_{x \in K} |\widehat{g}_\chi(x) - \delta_\chi(x)| \to_p 0$. Query feasibility, realized outcome observation, effective counts, and uncertainty remain separately visible in the finite-sample emission record.

Table 5: Episode and request ownership. The shared 5,000-replay subset is the matched-cost resource, not an additional pool per method.

Pool	Role	Separation
4,200 ToolBench+API-Bank episodes	outer-fold fitting, one update, post-update task evaluation	evaluation fold excluded from fit/tune/update
15,000 logged replays	5,000 shared by matched rows; full pool only where labeled	episode IDs inherit outer folds
612 adjudicated failures	417-request complete ToolBench+API-Bank slice for primary localization	labels excluded from every model stage
300 held-out requests	fixed-bank and blinded localization endpoints	disjoint from all fitting, tuning, and first audit

Table 6: Disjoint 300-request validation. Replay averages a fixed bank of three catalog draws per segment; blind uses three non-author annotators with earliest-minimal-repair tie breaking. Columns report emitted/all-request top-1 accuracy, cells use pooled counts, direction-agreement ECE is emission-only, and paired five-seed differences are reported in Section 6.

Method	Abstain	Replay emitted/all	Blind emitted/all	Dir.-ECE
Forced typed	0/300	149/300 (49.7/49.7%)	152/300 (50.7/50.7%)	0.081
Calibrated selective	28/300 (9.3%)	147/272 (54.0/49.0%)	144/272 (52.9/48.0%)	0.050
TRACEPACT	31/300 (10.3%)	164/269 (61.0/54.7%)	163/269 (60.6/54.3%)	0.043

Table 7: Catalog analysis with five-seed means and 95% intervals. Direction-agreement ECE uses emitted requests. Default rows reproduce the main operating point; held-out shifted rows compare forced emission, coverage-matched calibrated filtering, and the full support predicate.

Catalog	Method	Task	Emit. loc.	Abstain	Dir.-ECE
Conservative	TRACEPACT	71.0 ± 0.7	61.7 ± 1.4	5.5%	0.040
Default	TRACEPACT	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
Expanded	TRACEPACT	71.1 ± 0.7	60.3 ± 1.5	18.7%	0.047
Held-out shifted	Forced typed	69.1 ± 0.8	46.8 ± 1.8	0.0%	0.128
Held-out shifted	Calibrated selective	70.0 ± 0.8	52.1 ± 1.7	26.4%	0.081
Held-out shifted	TRACEPACT	70.7 ± 0.8	58.9 ± 1.6	26.4%	0.054

B CONTEXTUAL AND SYNTHETIC RESULTS

For each seed, the synthetic generator samples a 10–50-node acyclic graph whose nodes use the same four types as the real traces. A node state follows a type-specific linear structural equation in its parents plus independent Gaussian noise with $\sigma = 0.3$; terminal success thresholds a linear function of the terminal state. Each type-specific replacement catalog independently redraws that node's structural innovation and propagates the resulting state through the graph, so the catalog-average effect in Eq. 1 is available from the generator. Replay observability masks terminal outcomes as a function of observed typed context, arm, and replacement identity, using the same binary assignment rate as the real protocol. The hostile suite replaces the linear map with nonlinear parent interactions and heteroskedastic noise; its hidden-state slice removes a causal parent from x, and its contamination slice substitutes an invalid catalog draw for 30% of replacement requests. Oracle-answerable queries have a type-valid replacement and nonzero potential replay observability in both arms; known-unanswerable queries violate at least one of these conditions.

Table 8: Synthetic DAG evaluation on 5,000 source traces using the full 15,000-replay pool. RMSE and rank correlation use the common oracle-answerable set before gating; abstention uses all requests and direction-agreement ECE uses emitted records. Values are five-seed means with 95% CIs.

Method	Effect RMSE ↓	Rank corr. ↑	Abstain	Dir.-ECE ↓
Uniform	0.312 ± 0.011	0.00	0.0%	–
Saliency	0.198 ± 0.009	0.51 ± 0.03	0.0%	0.142
Shapley approximation	0.155 ± 0.008	0.64 ± 0.02	0.0%	0.098
Hindsight advantage	0.172 ± 0.010	0.58 ± 0.03	0.0%	0.121
Leave-one-out	0.141 ± 0.007	0.68 ± 0.02	0.0%	0.087
TRACEPACT	0.082 ± 0.005	0.83 ± 0.02	8.4%	0.031

On the common answerable set, TRACEPACT reduces RMSE by 47% relative to Shapley approximation and reaches 0.83 rank correlation; refusal therefore cannot improve the reported RMSE by removing difficult answerable queries. Figure 1 separates the larger-pool 0.082 result from the 5,000-replay frontier point at 0.085.

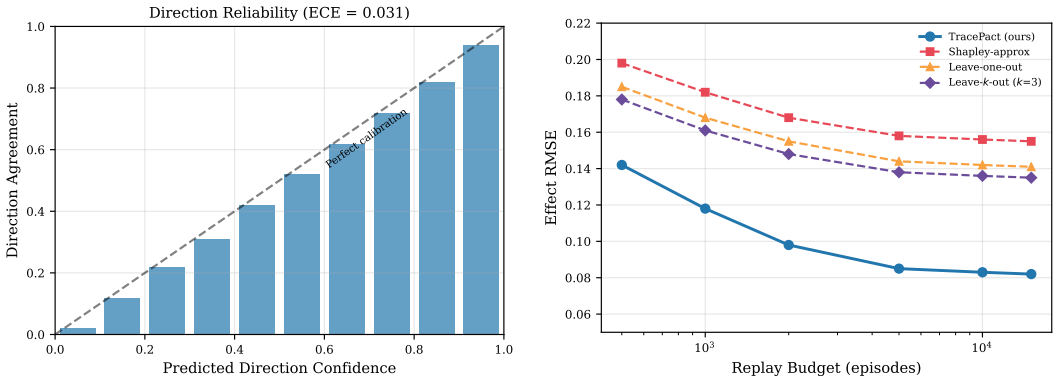


Figure 1: Five-seed synthetic diagnostics. Left: emitted direction confidence versus observed direction agreement at the Table 8 operating point (direction-agreement ECE $= 0.031$). Right: mean replay-cost frontier on the common oracle-answerable set; TRACEPACT's 5,000-replay RMSE is 0.085, while Table 8 reports 0.082 from the full 15,000-replay pool.

The direction-reliability bars track the diagonal across the emitted range, while the replay-cost frontier preserves TRACEPACT's ordering from 500 to 15,000 replays. These plots connect direction-agreement ECE to its replay-based sign target and the RMSE result to replay cost.

Table 9: ToolBench+API-Bank contextual comparison. Entries are five-seed means and 95% CIs, with direction-agreement ECE computed on emissions. The TRACEPACT row is exactly the operating point in Table 2, which provides the matched-cost comparison; the remaining rows show broader attribution context.

Method	Task success ↑	Emit. top-1 loc. ↑	Dir.-ECE ↓
Uniform weighting	67.6 ± 0.8	–	–
Saliency-weighted	68.4 ± 0.9	41.2 ± 1.8	0.142
Shapley-weighted	69.1 ± 0.7	48.7 ± 1.5	0.098
Hindsight-weighted	68.8 ± 0.8	45.3 ± 1.7	0.121
TRACEPACT-weighted	71.4 ± 0.6	62.8 ± 1.3	0.038

Across the broader suite, TRACEPACT raises task success over uniform weighting and improves both localization and direction calibration over saliency, Shapley, and hindsight weighting.

Table 10: WebArena-Trace results on 1,850 episodes with a matched 5,000-replay budget. Values are means and 95% CIs across five seeds; direction-agreement ECE is emission-only.

Method	Task success	Emit. loc.	Abstain	Dir.-ECE
Uniform weighting	38.2 ± 1.1	–	0.0%	–
Shapley-weighted	39.8 ± 1.0	39.4 ± 2.1	0.0%	0.112
Leave-k-out ($k = 3$)	40.3 ± 0.9	43.1 ± 1.8	0.0%	0.094
Off-policy advantage	40.1 ± 1.0	41.7 ± 1.9	0.0%	0.103
TRACEPACT-weighted	42.4 ± 0.8	54.6 ± 1.5	12.1%	0.042

On the longest-horizon domain, TRACEPACT improves task success by 2.1 points over matched leave-k-out, reaches 54.6% emitted localization, and emits at 0.042 direction-agreement ECE. The same record-level decision therefore remains useful on heterogeneous web interactions without changing the evaluation semantics.

C OPERATING-POINT STABILITY

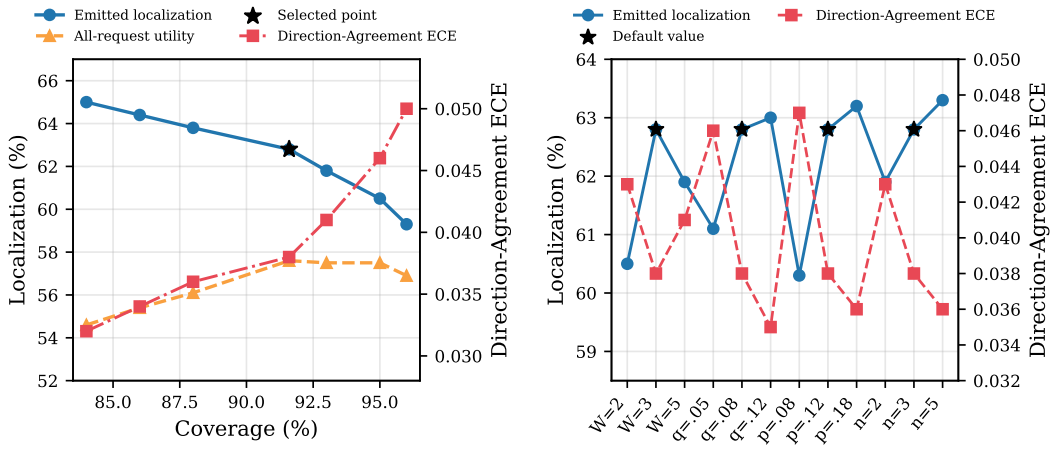


Figure 2: Operating-point stability on ToolBench+API-Bank. Left: uncertainty-threshold risk–coverage frontier, with the selected 91.6% point starred and a broad all-request utility plateau. Right: one-factor sweeps over replay window W, replay-observability floor ϵ_q, perturbation rate p_{perturb}, and effective-count floor.

The uncertainty frontier peaks in all-request utility around the selected point and remains flat through 95% coverage, while the one-factor sweeps preserve the same performance band across replay, observability, assignment, and count settings.

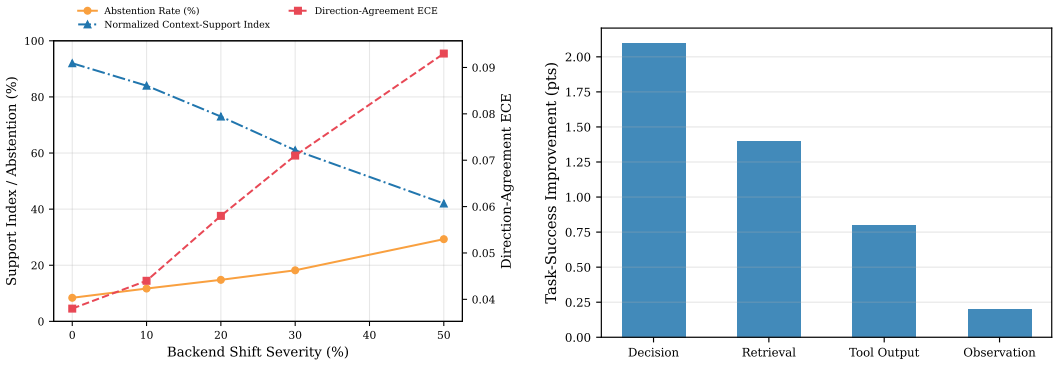


Figure 3: Support-shift and node-type diagnostics, averaged over five seeds. Left: normalized context-support index, abstention, and emission-only direction-agreement ECE under backend shift around the ToolBench+API-Bank operating point; randomized arm propensity remains fixed. Right: task-success points over all evaluation episodes from enabling each typed contribution in the common updater. Decision and retrieval dominate; observation adds $+0.2$.

Backend shift reduces contextual support and raises refusal before the randomized arm propensity changes. The node-type panel complements this reporting behavior by showing that decision and retrieval effects supply most of the downstream update gain.

Table 11: One-factor operating-point sweep on ToolBench+API-Bank. Every row uses five seeds and 5,000 replays; intervals are paired 95% bootstrap CIs. The $W = 3$, $\epsilon_q = 0.08$, $p_{\text{perturb}} = 0.12$, floor-3 row is shared.

Setting	Task success	Emit. loc.	Abstain	Dir.-ECE
$W = 2$	70.9 ± 0.7	60.5 ± 1.5	7.3%	0.043
$W = 3$	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
$W = 5$	71.2 ± 0.7	61.9 ± 1.4	10.2%	0.041
$\epsilon_q = 0.05$	71.1 ± 0.7	61.1 ± 1.5	5.8%	0.046
$\epsilon_q = 0.08$	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
$\epsilon_q = 0.12$	71.3 ± 0.6	63.0 ± 1.4	13.4%	0.035
$p_{\text{perturb}} = 0.08$	70.8 ± 0.7	60.3 ± 1.5	15.6%	0.047
$p_{\text{perturb}} = 0.12$	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
$p_{\text{perturb}} = 0.18$	71.5 ± 0.6	63.2 ± 1.3	5.4%	0.036
Count floor 2	71.2 ± 0.7	61.9 ± 1.4	5.9%	0.043
Count floor 3	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
Count floor 5	71.1 ± 0.7	63.3 ± 1.4	13.7%	0.036

No one-factor setting reverses the method's localization/calibration pattern; stricter observability and count thresholds exchange coverage for small direction-agreement ECE improvements.

Table 12: Five-seed mean uncertainty-gate frontier underlying Figure 2. Thresholds are selected on validation folds; all-request utility counts every refused request as incorrect, direction-agreement ECE uses emissions, and paired bootstrap intervals determine the stable plateau.

Coverage	Emit. loc.	All-request utility	Dir.-ECE
84.0%	65.0%	54.6%	0.032
86.0%	64.4%	55.4%	0.034
88.0%	63.8%	56.1%	0.036
91.6%	62.8%	57.6%	0.038
93.0%	61.8%	57.5%	0.041
95.0%	60.5%	57.5%	0.046
96.0%	59.3%	56.9%	0.050

All-request utility reaches 57.6% at 91.6% coverage and stays within 0.1 points through 95%, showing that the selected threshold lies on a broad plateau rather than a sharp optimum.

D Robustness and Transfer

A hostile synthetic suite replaces linear equations with nonlinear interactions and heteroskedastic noise, then separately injects hidden-state omission and 30% catalog contamination. Each condition uses five seeds and scores common-answerable-set RMSE, coverage, direction-agreement ECE, and false emissions on known-unanswerable queries.

Table 13: Hostile synthetic stress slices. Entries are five-seed means on the fixed nonlinear suite; direction-agreement ECE uses emissions, false emission uses known-unanswerable requests, RMSE uses the common oracle-answerable set, and coverage uses all requests.

Stressor	Metric	Calibrated selective	TracePact
30% catalog contamination	Direction-agreement ECE	0.096	0.066
30% catalog contamination	False emission	22.1%	9.4%
Severe hidden-state omission	Effect RMSE	0.151	0.148
Severe hidden-state omission	Coverage	82.0%	68.5%

Under catalog contamination, the certificate cuts false emission by 12.7 points and direction-agreement ECE by 0.030. Under severe hidden-state omission, TRACEPACT records 0.148 RMSE at 68.5% coverage, compared with 0.151 RMSE at 82.0% coverage for calibrated filtering. The reduced coverage is a visible response to incomplete conditioning while effect RMSE differences narrow.

Table 14: Transfer across acting-policy scale and downstream consumer on ToolBench+API-Bank. All rows use the common catalog, five seeds, and 5,000 replays; task and localization entries are means with 95% CIs, and direction-agreement ECE uses emissions.

Acting policy / consumer	Method	Task	Emit. loc.	Abstain	Dir.-ECE
Llama-3.1-70B / BC	Forced typed	70.1 ± 0.8	53.8 ± 1.6	0.0%	0.084
Llama-3.1-70B / BC	TRACEPACT	71.4 ± 0.6	62.8 ± 1.3	8.4%	0.038
Llama-3.1-8B / BC	Forced typed	62.0 ± 0.9	50.9 ± 1.7	0.0%	0.091
Llama-3.1-8B / BC	TRACEPACT	63.6 ± 0.8	59.4 ± 1.6	10.9%	0.043
Llama-3.1-70B / AWAC (Nair et al., 2020)	Forced typed	70.6 ± 0.7	54.5 ± 1.5	0.0%	0.080
Llama-3.1-70B / AWAC	TRACEPACT	72.2 ± 0.6	63.0 ± 1.3	8.7%	0.039

Both transfer blocks preserve the matched-cost ordering: the gain appears with a smaller acting policy and with AWAC, while direction-agreement ECE remains below the corresponding forced-typed value.

E CONTRACT SEMANTICS EVIDENCE MATRIX

The matrix below separates the evidence for query identity, support-conditioned emission, numerical effect estimation, and downstream use. Each layer has a dedicated test: catalog shifts test identity, support erosion tests emission, synthetic effects test estimation, and disjoint labels plus common updaters test decisions.

Table 15: Evidence map for the contract-indexed replay semantics. Each row pairs one claim with its validity control and the experiment that measures its consequence.

Layer	Claim and control	Evidence and consequence
Identity	δ_χ is indexed by type, catalog, W, and R; aligned contracts are the comparison control.	27.3% rank changes and 16.8% sign changes across catalogs; Eq. 1, Table 7.
Transition	Frozen snapshots and the terminal checker define continuation; observability heads report whether its endpoint is observed.	The support index and refusal respond to backend shift; Section 3.2, Figure 3.
Emission	Per-arm observability, effective counts, assignment, and uncertainty define the query rule.	Backend shift and cumulative gate ablation expose refusal behavior; Figure 3, Tables 4 and 12.
Estimation	Identity-aware cross-fitting estimates $\delta_\chi(x)$; validation calibration maps confidence to replay sign agreement.	Known-effect recovery and cumulative components separate estimator and calibration; Proposition 1, Table 8, Table 4.
Decision	Disjoint replay/blinded labels and a common updater test ranking and downstream use.	Concordant endpoint ordering and policy/updater transfer; Table 6, Eq. 4, Table 14.

The map keeps effect recovery distinct from repair localization: Eq. 1 is checked on generator-known effects, whereas Table 6 evaluates the decision induced by ranking repair lift.