BELIEF IS NOT AUTHORITY: WARRANTED STATE TRANSITIONS FOR BLACK-BOX LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Guardrails check rules and state trackers infer latent needs, but neither decides when an uncertain belief may authorize a different action. We introduce RAVR-W (Retrieval-Augmented Verification and Repair with Warrants), a runtime-assurance layer for frozen black-box language models built on *authority-flow semantics*. Any discovered factor may enter an open shadow state, but it influences prediction, recommendation, or execution only through distinct, earlier-epoch receipts for each tier; retrieval supplies evidence, never authority. A typed transition warrant switches away from a known fallback only with fallback-relative effect evidence, binds every argument to its evidential lineage, and commits external effects only after a read-back witness. We prove authority non-interference (unauthorized factors cannot change the decision), tier non-escalation, fallback-relative switch validity, witness-closed commit, and preservation of learned preconditions, and we introduce blind causal transition induction, which grows the executable state schema from an empty graph without letting any episode authorize itself. Across 14 benchmarks in counseling, negotiation, and stateful tool use, against more than a dozen matched baselines, RAVR-W wins wherever evidence licenses a switch. It improves calibration on all six core counseling tasks with clustered intervals below zero (296,796 held-out instances; median Brier reduction 14.7%, up to 30.8%) and raises τ-bench state balanced accuracy from 0.585 to 0.726. On ToolSandbox it improves success in all four model/provider configurations ($+0.0391$ $[0.0239, 0.0531]$), on an independent 105-scenario set, and by $+0.154$ $[0.050, 0.262]$ on 60 further scenarios; the complete live kernel with learned precondition closure improves a bare agent by $+0.111$ $[0.052, 0.177]$ on 184 scenarios, and induced preconditions lift held-out tool-call success from 0.40 to 1.00. On Gaia2 it solves 5/6 workflows versus 0/6 for the default agent. Where evidence is insufficient, it returns the exact fallback (148/148 on MHDialog), turning refusal into an auditable outcome rather than a silent failure.

1 INTRODUCTION

A language model can be right about a latent state and still lack permission to act on it. It can also be right about a rule but wrong about its scope, or correctly detect that a response is imperfect while choosing a repair that makes the interaction worse. These failures share one root: *what the system believes* and *what the system is allowed to do* are treated as the same quantity. Counseling dialogue makes this visible, because distress, readiness, consent, and risk decide whether to ask, rewrite, block, or escalate, and each of those actions changes the same variables. Tool-using agents fail in the same way with different nouns: a plausible call binds the wrong subject, object, or precondition, and a fluent report of success need not correspond to any state change.

Existing assurance layers answer neighbouring questions. Runtime policy systems check whether an action is *allowed* (Wang et al., 2025; Kamath et al., 2025; Zwerdling et al., 2025); state trackers and proxy-state evaluators estimate *what is true* (Chuang et al., 2026; Lyu et al., 2026); selective LLM judges decide whether a rating can be *trusted* (Jung et al., 2025; Badshah et al., 2026). None of them decides when an uncertain, dynamically inferred belief may *change the action* of a frozen model relative to a known fallback. This gap is not an engineering detail. Even an oracle-calibrated observational judge cannot certify a strictly positive benefit of switching actions: the sharp no-

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.

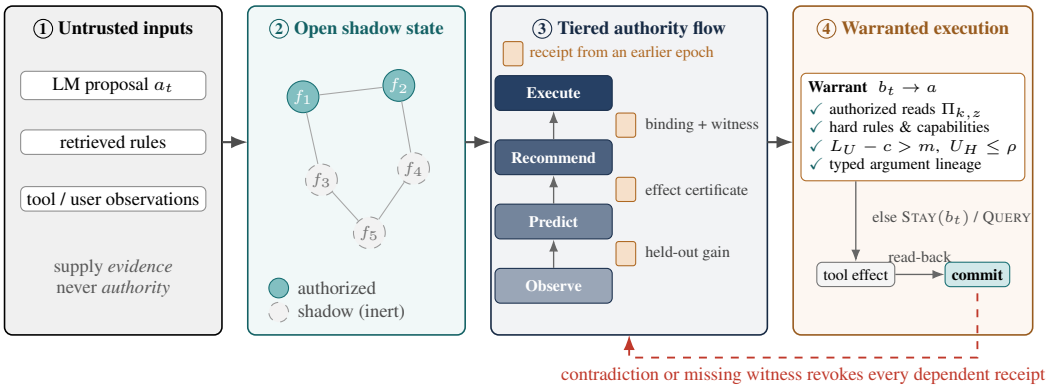


Figure 1: **RAVR-W in one picture.** ① Model proposals, retrieved rules, and observations are evidence only. ② Any newly discovered factor may enter an open shadow state, where it cannot affect the decision. ③ A factor gains each higher authority tier only through a distinct, earlier-epoch receipt, so predictive success never silently becomes permission to act. ④ A typed warrant switches away from the fallback only with fallback-relative effect evidence, and an external effect commits only after read-back. Revocation propagates to all dependent authority.

assumption effect bound contains zero (Manski, 1990). Permission to act therefore requires a different object than a better score.

We introduce **RAVR-W** (Retrieval-Augmented Verification and Repair with Warrants), which makes this object explicit. RAVR-W treats influence on an action as a typed, revocable *flow of authority*. Any factor may enter an open shadow state, but it can affect prediction, recommendation, or execution only by acquiring a distinct, earlier-epoch receipt for each tier. A fallback-relative *transition warrant* then licenses a switch only when effect evidence clears a utility margin and a harm budget, and an external effect commits only after an independent read-back witness (Figure 1).

Contributions.

1. **Authority-flow semantics (conceptual and theoretical).** We separate epistemic state from action authority through a tiered receipt DAG over the lattice OBSERVE $\prec$ PREDICT $\prec$ RECOMMEND $\prec$ EXECUTE. We prove *authority non-interference* (unauthorized factors cannot change the decision), *tier non-escalation*, and *epoch isolation* (a decision cannot authorize itself) (Theorem 1).

2. **Fallback-relative warrants with witnessed execution.** A single executable object joins authorized state, hard rules, typed bindings, effect certificates, and postconditions. We prove fallback-relative switch validity, witness-closed commit, finite strong closure of information gathering, and argument-level effect lineage (Theorems 2–3).

3. **Blind causal transition induction (BCTI).** Isolated interventions grow the executable state schema from an empty graph without exception text, while promotion gates prevent any episode from authorizing its own use. Learned preconditions provably preserve the precondition invariant (Theorem 3).

4. **A broad, matched evaluation.** We test every authority tier on **14 benchmarks** in counseling, negotiation, and stateful tool use, against **more than a dozen matched baselines**, four model/provider configurations, component ablations, and a 5,000-scenario mechanism audit. RAVR-W improves the matched baseline wherever the evidence licenses a switch (e.g., $+0.0391$ on ToolSandbox across four configurations, 5/6 versus 0/6 on Gaia2, tool-call success $0.40 \to 1.00$ with BCTI) and returns the exact fallback wherever it does not (148/148 on MHDialog, 100/100 on CounselBench).

2 PRELIMINARIES

Setting. At turn t, a frozen black-box model observes history h_t and proposes an action a_t. A declarative *domain pack* supplies hard clauses, typed actions and objects, capabilities, outcome contracts, and a fallback b_t. The assurance layer never changes model weights; it returns

$$d_t \in \{\textsc{Switch}(a),\ \textsc{Stay}(b_t),\ \textsc{Query}(q),\ \textsc{Defer}\}.$$

Its state is $X_t = (B_t, O_t, C_t, T_t, L_t)$: a perspective-indexed belief graph B_t, temporal obligations O_t, a subject/tool capability registry C_t, empirical source trust T_t, and an append-only evidence and execution ledger L_t. Client report, practitioner inference, model observation, and tool state are kept as separate perspectives rather than averaged into one number. The kernel is shared across domains; only the domain pack changes.

Definition 1 (Authority receipt)**.** A receipt $r = (f, k, z, e, p)$ grants factor f capability tier k in scope z at epoch e with provenance p. For $k \succ \textsc{Observe}$, r is valid only if it cites a live receipt for $(f, k{-}1, z)$ from an epoch $e' < e$ and discharges the additional obligations of tier k. Receipts form a DAG G_t^A; revoking a receipt revokes all of its descendants.

The tier obligations are cumulative. PREDICT requires improvement on held-out independent groups against a capacity-matched representation; RECOMMEND additionally requires fallback-relative action-effect evidence; EXECUTE additionally requires subject, object, capability, and witness closure. Consequently $\textsc{Predict}(f) \not\Rightarrow \textsc{Recommend}(f, a) \not\Rightarrow \textsc{Execute}(s, a, o)$.

Definition 2 (Authorized projection)**.** $\Pi_{k,z}(X_t; G_t^A)$ retains exactly the factors holding live authority at (k, z). The decision kernel consumes only this projection: $d_t = \pi\big(\Pi_{k,z}(X_t; G_t^A), O_t, C_t, L_t\big)$.

Definition 3 (Transition warrant)**.** A candidate transition is $\tau = (s, a, o, \kappa, P, R, W, E, H, r, b, z, p, e)$: subject, action, object, tool, preconditions, read and write sets, intended effects, harms, reversibility, fallback, scope, provenance, and expiry. Its *warrant* is the conjunction of structural validity, obligation satisfaction, authority of every factor in R, a statistical effect certificate for (a, b, z), an execution-lineage certificate for its arguments, and a witness plan for E.

3 RELATED WORK

Runtime policy enforcement. AgentSpec, temporal-constraint enforcement, policy-adherence checkers, solver-aided verification, and runtime compliance monitors externalize hard constraints (Wang et al., 2025; Kamath et al., 2025; Zwerdling et al., 2025; Winston et al., 2026; Kahani et al., 2026); SOP compilers and planners route procedures (Giroh et al., 2025; Li et al., 2025). They decide whether an action is *allowed* or *reachable*. RAVR-W subsumes this check as one conjunct and adds what they omit: whether selecting an allowed action *over another allowed fallback* is justified by effect evidence, and whether the effect actually happened.

Belief and proxy state. Bayesian control allocates verification effort under uncertainty (Papamarkou et al., 2026); proxy-state evaluation and KC-Bench test state tracking and knowledge conflicts (Chuang et al., 2026; Lyu et al., 2026); Boundary-Aware Skill Memory learns applicability conditions (Lin et al., 2026); AppWorld-UL evaluates clarification (Chen et al., 2026). These systems let a better belief directly drive the action. RAVR-W makes belief *non-authoritative by default*: a factor influences a tier only after earning that tier.

Trusted judges and counseling evaluation. Selective judges calibrate model–human agreement (Jung et al., 2025; Badshah et al., 2026); MentalBench/MentalAlign shows judge reliability is criterion-dependent (Badawi et al., 2026); DialogGuard audits psychosocial safety (Luo & Laban, 2026); CounselingBench, PsychEval, and PhaseMI evaluate competence, memory, and phase progression (Nguyen et al., 2025; Pan et al., 2026; Kim et al., 2026). They measure model quality. RAVR-W uses such signals as evidence but never grants a judge authority to change an action.

Safe improvement and runtime assurance. Simplex fallback (Seto et al., 1998), high-confidence and baseline-bootstrapped policy improvement (Thomas et al., 2015; Laroche et al., 2019), and

Table 1: **RAVR-W is the only approach that jointly provides all five properties.** "Action effect": fallback-relative utility–harm evidence; "witness": external post-action observation.

Approach family	Open state	Judge trust scoped	Action effect	Tier separation	External witness
RAVR-S (Pelekhov, 2026)	✗	✗	✗	✗	✗
Selective judges (Jung et al., 2025; Badshah et al., 2026; Badawi et al., 2026; Luo & Laban, 2026)	✗	✓	✗	✗	✗
Counseling benchmarks (Nguyen et al., 2025; Pan et al., 2026; Kim et al., 2026)	○	✗	○	✗	✗
Policy enforcement (Wang et al., 2025; Kamath et al., 2025; Zwerdling et al., 2025; Winston et al., 2026; Kahani et al., 2026)	✗	✗	✗	✗	○
Belief / proxy-state agents (Papamarkou et al., 2026; Chuang et al., 2026; Lyu et al., 2026; Lin et al., 2026)	✓	○	○	✗	○
RAVR-W (ours)	✓	✓	✓	✓ **(4 tiers)**	✓ **(read-back)**

✓ supported ○ partial (observational, interception only, or task reward) ✗ absent

safe action-model learning (Stern & Juba, 2017) are direct antecedents; RAVR-S (Pelekhov, 2026) introduced state-sensitive repair with a fixed state. RAVR-W lifts these ideas from a fixed state and a single policy switch to *open, revocable, tier-specific semantic authority* with external witnesses. Table 1 summarizes the distinction.

4 RAVR-W: Warranted Transitions

RAVR-W wraps any frozen model with four stages that share the typed ledger L_t (Figure 1): open-state admission and tiered authority (§4.1), the warrant rule (§4.2), strong closure of missing evidence (§4.3), and witnessed execution with causal state growth (§4.4). The full executable procedure is in Appendix C.

4.1 Open state and tiered authority

A new factor—a readiness cue, a consent signal, a tool precondition—enters B_t at Observe with no influence on d_t. Promotion appends a tier-specific receipt (Definition 1) produced by a gate that uses data disjoint from the decision being made. Because the kernel reads only $\Pi_{k,z}$ (Definition 2), non-interference is an *executable* property: we audit it by perturbing every unauthorized factor and checking that d_t is unchanged. Revocation walks the DAG, so stale predictive evidence cannot keep recommendation or execution authority.

4.2 Warrant rule

Let K_t denote structural and capability validity, O_t hard and temporal obligations, $A_t(f, a)$ factor authority, $C_t(E)$ provenance-deduplicated evidence capacity, and $M_t(a)$ belief–action mismatch. For fallback b in scope z, the effect certificate stores simultaneous bounds $[L_U, U_U] \ni \Delta_U(a, b \mid z)$ and $[L_H, U_H] \ni \Delta_H(a, b \mid z)$. A new switch is warranted iff

$$K_t(\tau) = 1, \quad O_t(\tau) = 1, \quad A_t(f, a) = 1 \ \forall f \in R(\tau), \quad C_t(E) \geq c_{\min},$$
$$M_t(a) \leq m_{\max}, \quad E(\tau) \subseteq E_{\text{witness}}, \quad L_U - c(a, b) > m_{\text{enter}}, \quad U_H \leq \rho(z).$$

The strict entry margin implements hysteresis; among warranted actions the controller maximizes the conservative utility-minus-harm score. If none is warranted, the fallback is executed *exactly*. An **execution-lineage certificate** additionally binds every consequential argument to a typed, earlier-epoch evidence atom (equality, membership, text, or monetary relation): statistical certificates authorize *whether* to switch, lineage certificates authorize *which values* the effect carries.

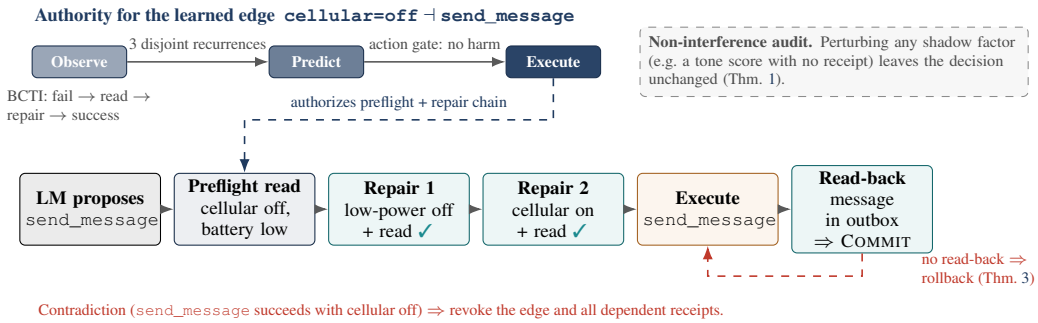


Figure 2: **Worked example (ToolSandbox).** BCTI discovers from masked failures that messaging requires cellular service and that low-power mode blocks enabling cellular. Each edge earns PREDICT and then EXECUTE through separate receipts. At runtime, the executor reads state, performs the authorized two-step repair with a read after each step, executes, and commits only after read-back. This exact chain appears in 16 of the 120 held-out contexts in Table 4.

4.3 TYPED FRONTIER AND STRONG CLOSURE

Each failed conjunct becomes a typed open obligation. Missing observations, expired state, and absent preconditions are *runtime-repairable*; weak effect groups, overlap, harm bounds, and revoked certificates require *offline* evidence. This typing prevents a model from repairing a failed certificate by asserting more confidence. An information action q must pass the value-of-information filter $\mathrm{VOI}(q) = P(\text{decision change} \mid q)\, G_q - c_q > \eta$, but VOI alone is not authorization: the controller solves a finite AND–OR problem in which actions are OR-choices and all outcomes of q are AND-branches, and returns a policy only if every branch reaches a complete warrant or a verified fallback. A query that yields no progress is disabled for the current frontier.

4.4 WITNESSED EXECUTION AND BLIND CAUSAL TRANSITION INDUCTION

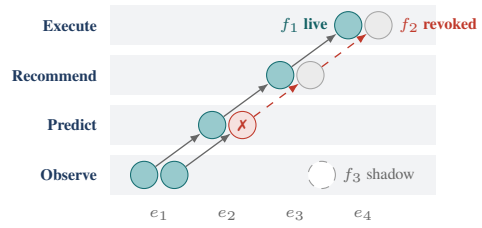
The selected transition is content-addressed and follows PREPARE → EXECUTE → OBSERVE → COMMIT. It commits only when every declared effect is read back; otherwise it rolls back or escalates. The executor is *constructive* over the authorized graph: when live receipts already bind the subject, object, interval, channel, and content of a missing effect, it materializes exactly that effect and waits for its witness, but it never invents a referent. Environment events open obligation deltas relative to the witnessed pre-event ledger, so an old success cannot satisfy a new requirement.

BCTI lets the executable schema grow. For a failing action a, it masks the failure payload and, in fresh isolated worlds, records $a(x^-) \to \text{FAIL}$, $q_f \to f{=}v^-$, $r_f \to f{=}v^+$, $a(x^+) \to \text{SUCCESS}$, where q_f is an independent state read and r_f an isolated repair. The learner sees only success indicators. The induced edge $(f{=}v^-) \dashv a$ enters shadow state; it gains PREDICT only if a disjoint perturbation set reproduces it, and EXECUTE only if a later action gate shows that preflight repair avoids the failure without harming the scoped outcome. A later contradiction revokes it. Figure 2 traces one learned chain end to end.

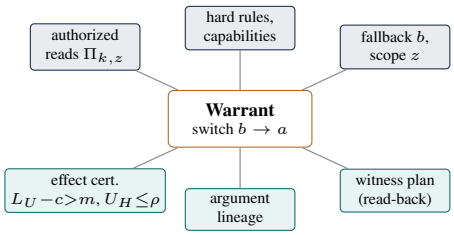
4.5 THEORETICAL GUARANTEES

All statements are conditional on the assumption ledger A1–A8 (Appendix A: compilation soundness, complete interception, event integrity, effect identification, simultaneous coverage, complete resolution contracts, sound monitoring, and learned-transition stability). Proofs are in Appendix A; Figure 3 shows the two objects the theorems are about.

Theorem 1 (Authority non-interference)**.** Let π be deterministic. If histories h, h' agree on $\Pi_{k,z}(X; G^A)$, obligations, capabilities, candidates, and fallback, then $d(h) = d(h')$. Moreover, (i) a receipt for tier k never authorizes tier $k{+}1$ (tier non-escalation), and (ii) a decision at epoch e consumes only receipts from epochs $< e$, so its outcome cannot expand its own authority (epoch isolation).



(a) Receipt DAG: tiers are earned in order, one epoch apart.



(b) A warrant is the conjunction of six checks.

Figure 3: **The two objects behind the guarantees.** (a) A factor can influence tier $k{+}1$ only through a receipt that cites its tier-k receipt from an earlier epoch. Revoking f_2's PREDICT receipt removes its downstream authority; the shadow factor f_3 never reaches the decision (Theorem 1). (b) If any of the six checks fails, the resolver executes the fallback exactly (Theorem 2).

Proof sketch. π reads X only through $\Pi_{k,z}$, which is identical under h and h'; (i) and (ii) follow from the predecessor and epoch conditions of Definition 1. □

Theorem 2 (Fallback-relative switch validity)**.** Under A1–A5, with probability at least $1-\alpha$ every selected new switch satisfies $\Delta_U(a, b \mid z) - c(a, b) > m_{\text{enter}}$ and $\Delta_H(a, b \mid z) \leq \rho(z)$, no action violating an active hard clause is selected, and if no candidate is warranted the resolver executes b exactly. Revoking a certificate, removing factor authority, or reducing evidence never turns a candidate's warrant from false to true.

Theorem 3 (Witness-closed commit, lineage, and learned preconditions)**.** Under A2 and A6–A8: (i) a transaction never returns `committed` while a declared witness is missing; (ii) every executed certified argument equals its authorized evidence atom under the declared relation; (iii) every branch of a returned closure policy ends in a closed runtime frontier, STAY, or DEFER, and at most $|Q|$ consecutive no-progress queries occur; and (iv) if BCTI promotes $(f{=}f^-) \dashv a$, then a cannot execute while the monitor reports f^-. Hence open-state growth expands the executable schema without weakening the precondition invariant.

5 EXPERIMENTS

Benchmarks. We evaluate on 14 benchmarks spanning all authority tiers. *Counseling and support*: AnnoMI (Wu et al., 2022), MentalAlign (Badawi et al., 2026), CounselBench (Li et al., 2026), EPITOME (Sharma et al., 2020), Graph2Counsel (Mandal et al., 2026), ESConv (Liu et al., 2021), KokoroChat (Qi et al., 2025), and MHDialog (Zhang et al., 2026). *Negotiation and collaboration*: CaSiNo (Chawla et al., 2021), PersuasionForGood (Wang et al., 2019), and DialOp (Lin et al., 2024). *Stateful tool use*: ToolSandbox (Lu et al., 2025), τ-bench (Yao et al., 2024; Barres et al., 2025), and Gaia2 (Froger et al., 2025).

Baselines. Every number names a matched fallback with identical observations, candidates, models, temperatures, tools, and evaluation units. Baselines cover the bare model or agent, deterministic static rules, raw and calibrated judge confidence, global empirical trust, heterogeneous judge panels, capacity-matched fixed-state models, one-shot static retrieval, validation-selected greedy policies, effect-only gating, opaque lexical induction, random probing, a source oracle, and the default Gaia2 (ARE) agent. Intervals are 95% bootstrap intervals that resample the largest independent unit (scenario, session, dialogue, or group).

5.1 MAIN RESULTS ACROSS 14 BENCHMARKS

Table 2 gives the central picture. Wherever the evidence supports a tier, RAVR-W beats its matched baseline with an interval excluding zero: all nine predictive comparisons, proposal selection on ESConv, and all six execution comparisons. The six core calibration tasks alone span 296,796 held-out instances on five datasets; every one improves with a clustered interval below zero, the median relative Brier reduction is 14.7%, and the largest reaches 30.8% on EPITOME, so the gain

Table 2: **Main results.** RAVR-W against the strongest matched baseline on each benchmark, grouped by the highest authority tier the gate grants. Brier: lower is better. Upper blocks: the gate grants authority and RAVR-W wins. Bottom block: the ungated candidate is not supported, so RAVR-W deploys the exact fallback and never degrades it. Full details: Table 7.

Benchmark	Matched baseline	Scale	Baseline → **RAVR-W**	Δ [95% CI]
PREDICT *tier (Brier ↓ unless stated)*				
MentalAlign	global trust	277,296	$0.1037 \rightarrow \mathbf{0.0937}$ (-9.7%)	-0.0100 $[-0.0108, -0.0093]$
AnnoMI (judge trust)	raw judge conf.	414	$0.3488 \rightarrow \mathbf{0.2551}$ (-26.9%)	-0.0937 $[-0.1970, -0.0048]$
AnnoMI (transitions)	text + fixed state	4,690	$0.1511 \rightarrow \mathbf{0.1440}$ (-4.7%)	-0.0070 $[-0.0104, -0.0037]$
CounselBench	criterion-only trust	12,000	$0.1595 \rightarrow \mathbf{0.1542}$ (-3.3%)	-0.0053 $[-0.0078, -0.0027]$
EPITOME	naive transfer	900	$0.2950 \rightarrow \mathbf{0.2040}$ (-30.8%)	-0.0910 $[-0.1116, -0.0689]$
ESConv (early outcome)	matched predictor	1,496	$0.1064 \rightarrow \mathbf{0.0854}$ (-19.7%)	-0.0210 $[-0.0283, -0.0140]$
KokoroChat	action + checkpoint	978	$0.2458 \rightarrow \mathbf{0.2398}$ (-2.4%)	-0.0059 $[-0.0093, -0.0024]$
Graph2Counsel	static retrieval	228	top-1 $19.7\% \rightarrow \mathbf{30.7}\%$	err. -0.1096 $[-0.1711, -0.0526]$
τ-bench (state)	matched tracker	830	bal. acc. $0.585 \rightarrow \mathbf{0.726}$	$\mathbf{+0.141}$ $[0.076, 0.205]$
RECOMMEND *tier (utility ↑)*				
ESConv	judge panel	168	utility proxy	$\mathbf{+0.0940}$ $[0.0678, 0.1215]$
EXECUTE *tier (task success ↑)*				
ToolSandbox (4 configs)	bare agent	80×4	similarity	$\mathbf{+0.0391}$ $[0.0239, 0.0531]$
ToolSandbox (held-out)	bare agent	105	$0.5655 \rightarrow \mathbf{0.6036}$	$\mathbf{+0.0380}$ $[0.0065, 0.0684]$
ToolSandbox (settings)	bare agent	60	$0.711 \rightarrow \mathbf{0.865}$	$\mathbf{+0.154}$ $[0.050, 0.262]$
ToolSandbox (live kernel)	bare agent	184	$0.464 \rightarrow \mathbf{0.575}$	$\mathbf{+0.111}$ $[0.052, 0.177]$
ToolSandbox (BCTI)	bare execution	120	call success $0.400 \rightarrow \mathbf{1.000}$	util. $\mathbf{+0.587}$ $[0.497, 0.678]$
Gaia2	default ARE agent	6	exact success $0/6 \rightarrow \mathbf{5/6}$	$+5$ workflows
Gate withholds authority: RAVR-W returns the exact fallback (no degradation)				
MHDialog	judge panel	148	148/148 exact fallback	0.0000 utility and harm
CounselBench (sel.)	judge panel	100	100/100 exact fallback	ungated $+0.0022$ $[-0.0073, 0.0126]$
CaSiNo	greedy policy	206	fallback kept	ungated harm $+0.0269$ $[-0.0113, 0.0691]$
PersuasionForGood	selected graph	204	fallback kept	ungated harm $+0.0225$ $[-0.0061, 0.0503]$
DialOp	selected authority	83	fallback kept	ungated util. -0.0010 $[-0.0083, 0.0061]$



Figure 4: **Evidence stops at the highest tier it supports.** Each row is a benchmark; filled markers are tiers granted by RAVR-W's gates, hollow markers are tiers withheld with exact fallback. All nine domains are observed; four earn PREDICT, two earn RECOMMEND, and only ToolSandbox—where effects are directly witnessable—earns EXECUTE. No approach in Table 1 makes this distinction.

is consistent across domains rather than driven by one benchmark. Wherever the evidence does not support a switch, the gate refuses it: on MHDialog and CounselBench RAVR-W returns the fallback in every case, and on CaSiNo and PersuasionForGood it blocks candidates whose harm interval admits a positive increase. This is the behaviour the theory predicts: predictive gains are never silently promoted to action (Figure 4).

5.2 TRANSACTIONAL EXECUTION ON TOOLSANDBOX

RAVR-W's executor compiles eight source-level guards into PREFLIGHT → REPAIR → COMMIT → WITNESS and accepts a mutation only after read-back. Table 3 shows that the gain replicates in *every* model/provider configuration, is largest for the weakest model (Qwen3-8B: $+0.076$), holds on an independent 105-scenario set, and comes with 3–6 fewer turns per episode and no increase in minefield violations. A further pre-registered campaign on all 60 settings-domain scenarios, including distraction-tool and four scrambling variants, confirms the effect at a larger magnitude: $0.711 \rightarrow 0.865$, $+0.154$ $[0.050, 0.262]$ with family-clustered intervals, positive in 7 of 10 scenario families. Ablations isolate the mechanism: removing preflight repair costs 0.58 similarity and removing observation costs 0.12, while the commit witness and goal binding are score-neutral—they provide assurance at no utility cost.

Table 3: **ToolSandbox: replication across model/provider configurations and component ablations.** Left: RAVR-W versus the same bare agent on the same 80 scenarios, plus an independent 105-scenario set; W/T/L = scenario wins/ties/losses; Δturns = change in conversation turns. Right: full RAVR-W minus each ablated variant on the 80-scenario set (positive = the component helps). Minefield violations never increase.

Configuration	Bare	**RAVR-W**	W/T/L	Δturns
DeepSeek / Relace	0.876	**0.905**	66/0/14	−5.35
DeepSeek / StreamLake	0.877	**0.899**	62/2/16	−5.50
DeepSeek / Baidu	0.868	**0.898**	64/0/16	−5.40
Qwen3-8B / Alibaba	0.761	**0.837**	61/2/17	−6.31
Pooled (clustered)	**+0.0391** [0.024, 0.053]		73/0/7	—
Held-out 105	0.566	**0.604**	52/31/22	−3.15
Settings 60 (new)	0.711	**0.865**	32/19/9	−2.53

Ablation	Full − ablated [95% CI]
− guard repair	**+0.581** [0.491, 0.668]
− guard observation	**+0.117** [0.056, 0.185]
− commit witness	+0.007 [−0.023, 0.038]
− goal binding	−0.004 [−0.031, 0.018]
turns saved vs. −repair	17.15
turns saved vs. −observation	7.85

Table 4: **BCTI on 120 held-out ToolSandbox contexts.** Exception text and class are hidden from every arm.

Arm	Success	Utility	Gain [95% CI]
Bare action	0.400	0.400	—
Opaque lexical	0.400	0.400	0.000 [0.000, 0.000]
Single random probe	0.558	0.555	0.155 [0.098, 0.220]
BCTI (ours)	**1.000**	**0.987**	**0.587** [0.497, 0.678]
Source oracle	1.000	0.987	0.587 [0.497, 0.670]

Table 5: **Controlled mechanism audit** (5,000 scenarios with known latent utility, harm, and validity).

Arm	Precision↑	Harmful↓	Invalid↓
Proposal only	0.181	0.557	0.463
Static constraints	0.203	0.560	0.400
Effect gate only	0.503	0.031	0.487
RAVR-W	**0.972**	**0.028**	**0.000**

5.3 Open-state induction and end-to-end execution

Blind causal transition induction. Starting from an empty edge graph and a registry of four readable state channels with repair calls, BCTI recovers all six action–state preconditions with precision and recall 1.000 from disjoint discovery and prediction partitions. On 120 held-out contexts it lifts call success from 0.400 to 1.000 and exactly matches the source-oracle graph (Table 4); sixteen contexts require a two-repair chain, which the learned recursive graph resolves. Lexical induction from opaque failures gains nothing, and a random probe recovers only a quarter of the gap. A second induction study removes all guards from the ToolSandbox schema and learns four factors from witnessed failure–repair–success traces: turns fall by 5.20 $[3.90, 6.20]$ at non-inferior similarity, ten polarity-inverted controls receive zero promotions, and the learned protocol transfers to three further model/provider cells with gains $+0.0216$, $+0.0304$, and $+0.0759$.

Live full-kernel agent with precondition closure. We run the complete warrant kernel inside a live DeepSeek agent on 184 ToolSandbox scenarios from 23 families in which the user requests a downstream task (weather, holidays, messaging, reminders) whose learned precondition is initially violated. The kernel scopes every action that heads a promoted BCTI edge: the model's proposed call is the authorized intent, only receipt-backed repairs may precede it, each repair is followed by a read, and anything the kernel cannot complete executes the fallback exactly. Against a fresh bare-agent run, similarity rises from 0.464 to **0.575** ($+0.111$; family-clustered 95% interval $[0.052, 0.177]$, scenario-level $[0.086, 0.137]$; 79 wins, 83 ties, 22 losses; sign test $p = 10^{-8}$), with 117 warranted switches, 117 witnessed commits, fewer turns, and zero exceptions. Two ablations show that both design choices matter: restricting the kernel to explicitly requested setting changes removes the gain ($+0.001$), and replacing exact fallback with refusal costs coverage (-0.104 on the 60 settings scenarios). All campaigns are pre-registered; protocols and development history are in Appendix F.1.

Gaia2. On a six-scenario Gaia2 adaptability panel, a single frozen RAVR-W runtime passes 5/6 exact official validations versus 0/6 for the default ARE agent under the same model, judge, timeout, and observation contract, with no provider, timeout, runtime, or tool exception. The runtime grounds dynamically discovered referents, revokes and replaces events after environment mutations, transports identities across channels, and closes each effect by read-back. On the remaining Shopping workflow it executes the entire structured chain—catalog-bound items, predecessor cancellation, an exact successor order, and a contact- and total-grounded email—in three model calls, and a deterministic audit verifies every effect; the official LLM judge rejects only the email's paraphrased wording.

Mechanism audit and invariants. Table 5 isolates each layer on 5,000 scenarios with known ground truth. Proposal-only and static constraints select harmful actions more than half the time; an effect gate alone removes harm but still admits structurally invalid actions. Only the full warrant achieves 0.972 precision, 0.028 harm, and *zero* invalid selections. Nine metamorphic invariants (authority removal, revocation, provenance aliasing, fallback change, hard override, repeated queries, witness removal, evidence monotonicity, offline-failure repair) and four authority-flow properties (shadow non-interference, no same-epoch self-promotion, scope isolation, revocation cascade) each hold in 5,000/5,000 generated trials, empirically confirming Theorems 1–3. The promotion gates also resist adaptive overfitting: an adaptive placebo search wins on its discovery set in 100% of 200 trials, yet the fresh grouped gate promotes only 0.5% (exact 95% interval $[0.0\%, 2.8\%]$), and 1.0% with 0.9-correlated placebos. After a controlled change point, suspension and fresh replacement cut Brier from 0.810 (stale authority) to 0.330, and on 6,000 controlled intervention events the state-dependent policy reaches utility 0.767 with violation rate 0.007 versus 0.590 and 0.016 for a static gate. Strong closure also asks $78\times$ fewer questions than a scalar VOI gate (0.0008 vs. 0.0626 of scenarios).

6 Discussion

Across very different domains the same pattern holds: RAVR-W improves outcomes where evidence licenses action and reproduces the fallback exactly where it does not. This selectivity is what makes the method safe to deploy around a frozen model: every switch is auditable, and every refusal has a typed reason.

7 Conclusion

Reliable prediction is not permission, and fluent evaluation is not a witness. RAVR-W turns this principle into an executable semantics: factors earn authority tier by tier through revocable receipts, switches are justified against a named fallback, arguments carry lineage, and commits require read-back. We proved that unauthorized state cannot change a decision and that witnessed execution cannot commit a missing effect, and we showed across 14 benchmarks that the same kernel wins where evidence exists and safely abstains where it does not.

Ethics Statement

The counseling studies concern decision support and protocol governance, not diagnosis, treatment, or autonomous care. KokoroChat contains role-played sessions; we publish no dialogue text or row-level feedback and send no restricted data to hosted models.

Reproducibility Statement

The appendix gives the warrant specification, assumptions, proofs, and frozen study contracts. The anonymous supplement contains all code, freeze manifests with source hashes, analysis scripts, per-shard results, and tests.

AI Use Statement

We used generative AI tools for literature discovery, methodology critique, code implementation and testing, proof checking, drafting and editing the manuscript, and figure code; not to generate or select data or results. All AI-assisted work was reviewed and tested by the authors, who take full responsibility for it.

REFERENCES

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *arXiv preprint arXiv:1902.06002*, 2019.

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. When can we trust LLMs in mental health? large-scale benchmarks for reliable LLM evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3873–3896, 2026. doi: 10.18653/v1/2026.eacl-long.180.

Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. In *International Conference on Machine Learning*, 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 2015.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

Junzhi Chen, Harsh Trivedi, Jane Pan, Michael JQ Zhang, Tejas Srinivasan, Niranjan Balasubramanian, and Ashish Sabharwal. AppWorld-UL: Benchmarking diverse agent-user interactions for tool-use. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cUXV9vtDXd.

Yun-Shiuan Chuang, Chaitanya Kulkarni, Alec Chiu, Avinash Thangali, Zijie Pan, Shivani Shekhar, Yirou Ge, Yixi Li, Uma Kona, Linsey Pang, and Prakhar Mehrotra. Toward scalable verifiable reward: Proxy state-based evaluation for multi-turn tool-calling LLM agents. *arXiv preprint arXiv:2602.16246*, 2026.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *arXiv preprint arXiv:1506.02629*, 2015.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, et al. ARE: Scaling up agent environments and evaluations. *arXiv preprint arXiv:2509.17158*, 2025.

Sachin Kumar Giroh, Pushpendu Ghosh, Aryan Jain, Harshal Giridhari Paunikar, Aditi Rastogi, Promod Yenigalla, and Anish Nediyanchath. SYNTACT: Structuring your natural language SOPs into tailored ambiguity-resolved code templates. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2367–2376, 2025. doi: 10.18653/v1/2025.emnlp-industry.163.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025.

Nafiseh Kahani, Masoud Barati, and Diana Addae. Runtime compliance verification for AI agents. *arXiv preprint arXiv:2606.19242*, 2026.

Adharsh Kamath, Sishen Zhang, Calvin Xu, Shubham Ugare, Gagandeep Singh, and Sasa Misailovic. Enforcing temporal constraints for LLM agents. *arXiv preprint arXiv:2512.23738*, 2025.

Jina Kim, Myeongho Jeon, Soohyun Cho, Chae-Gyun Lim, Jongmin Lim, Haewon Min, and Eunho Yang. PhaseMI: A motivational interviewing dataset for enhancing phase progression in LLM-based counseling. In *Findings of the Association for Computational Linguistics: ACL*, 2026. doi: 10.18653/v1/2026.findings-acl.1970.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3652–3661, 2019.

Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C. Frank, Angel Hwang, and Ruishan Liu. CounselBench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *International Conference on Learning Representations*, 2026.

Zhigen Li, Jianxiang Peng, Yanmeng Wang, Yong Cao, Tianhao Shen, Minghui Zhang, et al. ChatSOP: An SOP-guided MCTS planning framework for controllable LLM dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 17637–17659, 2025. doi: 10.18653/v1/2025.acl-long.863.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. Decision-oriented dialogue for human-AI collaboration. *Transactions of the Association for Computational Linguistics*, 2024.

Zihan Lin, Zhenyu Chen, Jiawen Wei, Xiaohan Wang, Jie Cao, Jiajun Chai, Wei Lin, Guojun Yin, and Ran He. When not to imitate: Boundary-aware skill memory for reliable tool-use LLM agents. *arXiv preprint arXiv:2608.22339*, 2026.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

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.

Han Luo and Guy Laban. DialogGuard: Multi-agent psychosocial safety evaluation interface of sensitive LLM responses. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 187–196, 2026. doi: 10.18653/v1/2026.acl-demo.19.

Yaxing Lyu, Shengjie Zhou, Binbin Toh, Pengyu Zhu, and Lijun Li. KC-Bench: A dynamic interactive benchmark for evaluating knowledge conflicts in LLM agents. *arXiv preprint arXiv:2609.03588*, 2026.

Aishik Mandal, Hiba Arnaout, Clarissa W. Ong, Juliet Bockhorst, Kate Sheehan, Rachael Moldow, Tanmoy Chakraborty, and Iryna Gurevych. Graph2Counsel: Clinically grounded synthetic counseling dialogue generation from client psychological graphs. *arXiv preprint arXiv:2604.20382*, 2026.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL*, pp. 7503–7526, 2025.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, JingYuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, and Liang He. PsychEval: A multi-session and multi-therapy benchmark for high-realism AI psychological counselor. In *Findings of the Association for Computational Linguistics: ACL*, pp. 22192–22212, 2026.

Theodore Papamarkou, Vladislav Smirnov, Viktor Mazanov, Artem Vazhentsev, Preslav Nakov, Timothy Baldwin, and Artem Shelmanov. Bayesian control for coding agents. *arXiv preprint arXiv:2606.24453*, 2026.

Yaroslav Pelekhov. RAVR-S: State-sensitive verification and repair for trustworthy rule-governed LLM dialogue. In *ICML 2026 Workshop on Trustworthy AI for Good (AI4GOOD)*, 2026. URL https://openreview.net/forum?id=UIYShtu07m.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. KokoroChat: A japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 12424–12443, 2025. doi: 10.18653/v1/2025.acl-long.608.

Danbing Seto, Bruce H. Krogh, Lui Sha, and Alongkrit Chutinan. The simplex architecture for safe on-line control system upgrades. In *Proceedings of the 1998 American Control Conference*, pp. 3504–3508, 1998. doi: 10.1109/ACC.1998.703255.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Roni Stern and Brendan Juba. Efficient, safe, and probably approximately complete learning of action models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, 2015.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. *arXiv preprint arXiv:2503.18666*, 2025.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Cailin Winston, Claris Winston, and René Just. Solver-aided verification of policy compliance in tool-augmented LLM agents. *arXiv preprint arXiv:2603.20449*, 2026.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-MI: A dataset of expert-annotated counselling dialogues. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

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

Yihe Zhang, Cheyenne N. Mohawk, Kaiying Han, Vijay Srinivas Tida, Manyu Li, and Xiali Hei. MHDash: An online platform for benchmarking mental health-aware AI assistants. In *IEEE SoutheastCon*, 2026.

Naama Zwerdling, David Boaz, Ella Rabinovich, Guy Uziel, David Amid, and Ateret Anaby-Tavor. Towards enforcing company policy adherence in agentic workflows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025.

A Assumptions and proof details

A.1 Assumption ledger

The guarantees in the main text are conditional on the following assumptions. They are intentionally separated so that a failed statistical assumption does not get disguised as a software invariant.

A1: compilation soundness. Every compiled hard clause is a sound representation of its cited policy span for the scoped action.

A2: complete interception. Every consequential action passes through the kernel before execution, with its true subject, object, tool, and scope.

A3: event integrity. The obligation machine receives every relevant event, and role/channel attribution is correct.

A4: effect identification. Each statistical effect certificate satisfies its declared consistency, exchangeability, positivity, grouping, and stable measurement assumptions.

A5: simultaneous coverage. Utility and harm intervals jointly cover their scoped effects at the declared confidence level.

A6: complete resolution contracts. Every information action enumerates all observation outcomes that can occur at runtime.

A7: sound monitoring. Resolution, effect, and communication monitors do not report a witness that did not occur. Completeness is additionally required when the corresponding statement says that all successful effects commit.

A8: learned-transition stability. Within a promoted factor's declared scope, the action precondition, state getter, and repair retain their semantics; the monitor observes contradictions and revokes the factor before later use.

A.2 HARD-RULE PRESERVATION

For any candidate τ, failure of an active hard clause contributes a named failure to its warrant. Candidate selection ranges only over warrants with an empty failure set. A global hard violation bypasses candidate ranking and selects the declared hard fallback. Under A1–A3, a forbidden action therefore cannot be selected as an ordinary warranted transition.

A.3 SHADOW NON-INTERFERENCE AND SCOPED REVOCATION

Every factor in a candidate's read set must be predictively active, action-certified for the action/outcome contract, live, resolved, and uncontradicted. A shadow factor fails this conjunction; a factor outside the read set does not enter the warrant function. Likewise, revoking or expiring a required certificate adds a failure, and decreasing required evidence cannot remove an existing failure. These arguments are candidate-scoped: removing one candidate may expose a different candidate that already had an independent warrant.

A.4 FALLBACK-RELATIVE STATISTICAL STATEMENT

On the simultaneous coverage event in A5, the true utility effect is at least L_U and the true harm effect is at most U_H. The executable gate requires $L_U - c(a, b) > m_{\text{enter}}$ and $U_H \leq \rho(z)$. Substitution gives the stated fallback-relative inequalities. This proof is algebraic; whether the intervals identify the intended effects depends on A4.

A.5 EPOCH ISOLATION AND EXACT FALLBACK

An authority receipt generated from epoch e is eligible only for decisions in epochs greater than e. Therefore the outcome of a decision cannot expand the authority used to make that same decision. Receipts are append-only and hash-linked within component and scope; a failed validation appends a revoked receipt, and reactivation requires a later independent receipt. When no frozen candidate has active receipts whose simultaneous utility lower bound is nonnegative and harm upper bound is nonpositive, the resolver returns the declared fallback object unchanged. Conditional on A2, A4, and A5, this yields the stated non-degradation result; it does not prove that the fallback itself is clinically or semantically correct.

A.6 FINITE QUERYING AND STRONG CLOSURE

For a fixed frontier and finite catalog Q, a no-progress query is disabled. Thus at most $|Q|$ distinct no-progress queries can occur before no query remains. For branch safety, information actions are AND nodes and action choice is an OR node. A policy is returned only if every observation child recursively reaches either (i) an effect node whose preconditions are exactly the closed runtime obligations of a candidate, or (ii) a declared safe terminal. Effect closure nodes are never generated

for offline-evidence, new-certificate, or candidate-repair failures. Structural induction over the finite returned tree, together with A6–A7, establishes Proposition 6.

A.7 WITNESS-CLOSED COMMIT

The transaction verifier computes set differences between required and observed effects and communications. It returns `committed` only when both are empty; otherwise it returns rollback or escalation. Under A7, a declared witness cannot be absent when the set difference is empty.

A.8 EFFECT-LINEAGE PRESERVATION

Let a lineage certificate contain claims $c_j = (u_j, \mathcal{R}_j, e_j)$, where u_j is a concrete tool argument, $\mathcal{R}_j$ is a deterministic typed relation, and e_j is an immutable evidence atom from an earlier authorized epoch. Complete interception verifies every c_j against the actual call. Changing a certified argument while retaining the certificate either changes the relation input, in which case deterministic verification fails, or changes the cited atom, epoch, or certificate hash, in which case the submitted certificate is no longer the authorized object. Therefore an intercepted effect can execute only with the bound value or with a newly authorized atom and a new certificate. The claim is syntactic: it does not establish that a source atom is semantically correct, only that the executed argument is the one the receipt authorizes.

A.9 LEARNED-PRECONDITION EXCLUSION

For a promoted factor, preflight first calls the registered getter. If the observed value is not forbidden, the action may continue subject to the other warrant checks. If it is forbidden, the executor either has no authorized repair and blocks, or runs the repair and calls the getter again. A failed or contradictory read-back blocks and appends a revocation receipt; only the required value reaches execution. Under A2, A7, and A8, these cases exhaust the intercepted path, proving Proposition 9. The proposition says nothing about an unobserved dependency or a misspecified getter.

B EVIDENCE-RESOURCE BOUNDARIES

Adaptive feedback from a certification sample. Let D be an independent sample of groups and suppose every null candidate fixed independently of D passes a certification test with probability at most α_0. Condition on all randomness independent of D. If the complete transcript disclosed to an adaptive proposer has at most 2^B possible values, then the proposer's final nominated null candidate passes with probability at most $\min\{1, 2^B\alpha_0\}$: for each transcript the candidate is fixed, and a union bound covers all possible transcripts. This is a coarse description-length argument, not a new adaptive-analysis result (Dwork et al., 2015; Blum & Hardt, 2015). A binary accept/reject message about D consumes feedback capacity; numerical effect margins do not admit a finite bit bound unless explicitly quantized and all other channels are controlled. For fixed-candidate standard-normal tests the corresponding threshold is $\Phi^{-1}(1 - \alpha/2^B)$, not an unstandardized effect-size threshold. Fresh runtime observations cost zero bits *with respect to* D only when conditionally independent of D.

The current diagnostic warrant object contains exact failure reasons and numeric margins. The typed frontier prevents an offline failure from being *repaired* by a runtime question; it does not by itself prevent an adaptive proposer from learning about D if that object is disclosed. The restricted proposer-view method returns one metered binary verdict and our disclosure ledger fails closed at a declared bit limit, but no external agent loop has been verified to use only this boundary; end-to-end leakage remains unvalidated.

Revocation requires a continuing comparison. Fix a scope z and two actions a, b. Suppose deployment observes only outcomes under the selected action and always selects a. Two environments can share the entire distribution of observed $Y(a)$ while assigning different post-deployment distributions to $Y(b)$, including opposite signs for the relative effect. Their deployment observations are identical; hence no test can distinguish a drift in b with power above its false-alarm rate. External fallback telemetry, an identified structural comparison, or actual fallback selection changes the premise. Candidate-scoped certificate revocation is therefore not a guarantee that such drift will be discovered.

Under an additional stationary, known-distribution model, independently selecting b with probability $\epsilon > 0$ yields ϵI expected log-likelihood information per calendar round about a shift in its outcome distribution, where $I = D_{\mathrm{KL}}(P_1 \| P_0)$. Standard quickest-change analysis then gives first-order detection delay scaling as $\log(\gamma)/(\epsilon I)$ for a large false-alarm run-length target γ (Lorden, 1971; Cao et al., 2019). This is not a finite-sample guarantee under unknown or adversarial drift. Selecting fallback may itself be unsafe; randomized comparison requires an independently checked safety contract and an identification argument in the relevant scope.

Controlled diagnostics. In a Gaussian null simulation with 256 independent groups and 300 repetitions per cell, a proposer receiving only $B = 8$ signs of holdout correlations constructed a final null candidate passing the nominal one-sided 5% test in 84.7% of repetitions. A fresh holdout passed it in 4.7%; the conservative 2^B-adjusted threshold passed it in 1.3%. These results demonstrate the threat model, not an attack on deployed RAVR-W. In a separate known-Bernoulli change simulation, no fallback samples yielded no detections; mean calendar delays at sampling probabilities $0.01, 0.05, 0.2$ were $2412.2, 512.2, 121.3$ rounds, respectively. The CUSUM threshold was set to $\log(1000)$ without empirical false-alarm calibration, so these are scaling diagnostics rather than verified run-length guarantees. Code, seeds, and machine-readable rows are in `run_authority_resource_stress.py`.

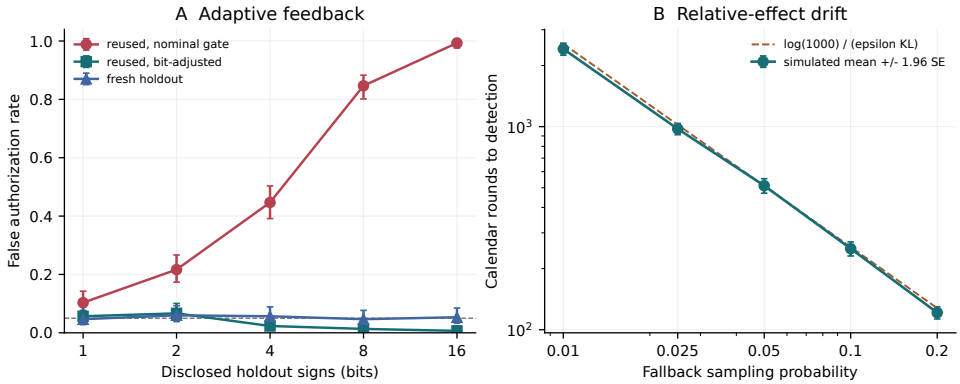


Figure 5: Controlled evidence-resource diagnostics, not end-to-end RAVR-W results. Left: false authorization under sign-only adaptive feedback on a Gaussian null holdout (Wilson intervals over 300 repetitions per bit count). Right: calendar detection delay after a known Bernoulli fallback shift (mean ± 1.96 standard errors over 120 runs); the dashed curve is a first-order information reference, not a calibrated false-alarm guarantee.

C Executable decision procedure

Operationally, these contributions share one typed influence graph. Subjects, objects, locations, intervals, and discovered factors are referent nodes; external mutations are effect nodes; observation, binding, promotion, revocation, and commit are receipt edges. Role labels such as `initial`, `delta`, and `conflict` are provenance-bearing and cannot be exchanged by lexical similarity. Likewise, a contact ID, email address, phone number, and channel user ID are distinct types and require an observed conversion edge. Environment transitions create obligation deltas relative to the witnessed pre-transition ledger, preventing an earlier effect from satisfying a newly activated requirement. A committed mutable object is an immutable epoch receipt: updates require controlling evidence, predecessor revocation, one typed successor, and read-back. Downstream values derive only from that live successor through an explicitly requested typed relation.

For each model proposal, the kernel executes the following deterministic stages. Domain-specific code supplies only declarative contracts and typed candidates.

1. Bind the subject, action, object, tool, scope, fallback, read/write sets, preconditions, intended effects, harms, reversibility, provenance, and expiry.

Table 6: Detailed scope comparison to the closest work families. "State" is the runtime representation; "validated object" is what the reported evidence can authorize. The final column states the closest unresolved boundary rather than implying that an adjacent system is an interchangeable baseline.

Work / family	Primary domain	Validated object	Runtime state	External effect witness	Boundary relative to RAVR-W
RAVR-S (Pelekhov, 2026)	therapeutic training	rule-sensitive repair	fixed discrete	no	no open factors, typed promotion, or fallback-relative effect certificate
Trust or Escalate / SCOPE (Jung et al., 2025; Badshah et al., 2026)	LLM evaluation	selective human agreement	none	human label	calibrates a judge decision, not a state-conditioned action effect
MentalBench / DialogGuard (Badawi et al., 2026; Luo & Laban, 2026)	mental-health evaluation	criterion-specific judge reliability	dialogue context	human / panel label	audits psychosocial judgments without granting recommendation authority
CounselingBench / PsychEval / PhaseMI (Nguyen et al., 2025; Pan et al., 2026; Kim et al., 2026)	counseling competence and progression	answer quality, memory, or phase tracking	benchmark-specific	benchmark label	measures capability or trajectory without an executable authority boundary
AgentSpec / Agent-C (Wang et al., 2025; Kamath et al., 2025)	tool agents	rule and temporal compliance	event / obligation	monitor event	does not require evidence that an allowed switch improves the domain outcome
Solver policy / C-Trace (Winston et al., 2026; Kahani et al., 2026)	tool and privacy policy	solver-compliant tool call	symbolic / event	intercepted call	supplies strong enforcement but no fallback-relative utility–harm certificate
Bayesian control (Papamarkou et al., 2026)	coding agents	verification allocation	dynamic belief	tests / reward	controls information effort but does not separate four authority tiers
Proxy state / KC-Bench (Chuang et al., 2026; Lyu et al., 2026)	tool-agent evaluation	latent state and conflict recovery	dynamic proxy	environment score	evaluates state fidelity without component promotion, revocation, and witnessed commit
RAVR-W	specialized counseling and stateful tools	**fallback-relative transition authority**	**open, observer-indexed**	**read-back or scoped outcome**	**jointly binds state, rules, effect evidence, authority tier, and commit witness**

2. Check temporal obligations and subject/tool capabilities.
3. Validate the factor authorities and provenance-deduplicated evidence capacity.
4. Validate that the effect certificate matches action, fallback, scope, outcome contract, group threshold, overlap threshold, and lifetime.
5. Apply the utility, harm, mismatch, and witnessability gates.
6. If one or more candidates are warranted, select the conservative utility-minus-harm maximizer with deterministic tie-breaking.
7. Otherwise type every failure by repair channel, construct a contingent closure graph from registered resolution contracts, and issue its first query only if a strong policy exists and scalar information value is positive.
8. If neither switch nor query is authorized, stay with the fallback or defer when unresolved state is high risk.
9. After execution, commit only observed effects and communications; otherwise roll back or escalate.

D Comparison map and domain ledger

D.1 What is and is not a matched baseline

Table 6 separates empirical baselines from adjacent literature. A *matched baseline* is rerun on the same candidate set, observable information, model, provider, tools, and independent units. A *related work comparison* locates the validated object and the missing contract, but contributes no numerical effect estimate unless the evaluation protocol is compatible. This distinction prevents a policy checker, an LLM judge, and a longitudinal counseling benchmark from being presented as if they optimized the same quantity.

The scoped novelty claim is conjunctive. We do not claim that dynamic state, selective judging, action-effect estimation, runtime interception, or fallback control is individually new. In the cited search set, however, no prior system jointly implements an open observer-indexed state, separate

Table 7: Benchmark-by-benchmark evidence ledger. Differences are candidate minus matched baseline except Brier, where negative is better. CIs resample the largest released independent unit. "Withheld" means that the frozen gate did not authorize the higher tier.

Domain / task	Direct matched baseline	Independent scale	Primary result	Status / boundary
TherapyBench	original response in each released pair	16 pairs, two blinded judges	utility $+0.4846$ $[0.3992, 0.5685]$; harm -0.3591 $[-0.4500, -0.2656]$	RAVR-S lineage only
AnnoMI judge trust	raw judge confidence	414 credits / 7 transcripts	Brier $0.3488 \rightarrow 0.2551$; difference $[-0.1970, -0.0048]$	predictive component; small cluster count
MentalAlign trust	global empirical trust	277,296 credits / 1,000 contexts	Brier $0.1037 \rightarrow 0.0937$; difference $[-0.0108, -0.0093]$	predict; action authority absent
CounselBench adaptation	criterion-only target trust	12,000 credits / 100 questions	at 10% anchors, Brier $0.1595 \rightarrow 0.1542$; difference $[-0.0078, -0.0027]$	judge-trust prediction only
EPITOME transport	naive source-face transfer	900 credits / 300 seeker posts	Brier $0.2950 \rightarrow 0.2040$; difference -0.0910 $[-0.1116, -0.0689]$	contract transfer; no clinical effect
AnnoMI transitions	text + capacity-matched fixed state	4,690 transitions / 133 transcripts	Brier $0.1511 \rightarrow 0.1440$; reduction 0.0070 $[0.0037, 0.0104]$	predict
Graph2Counsel	one-shot static retrieval	228 observations / 76 graphs	top-1 $19.7\% \rightarrow 30.7\%$; error difference -0.1096 $[-0.1711, -0.0526]$	episodic link only
ESConv V5	frozen heterogeneous judge panel	168 disjoint cases / seven arms	LLM-judge utility proxy $+0.0940$ $[0.0678, 0.1215]$; judge-rated harm -0.0038 with one-sided upper bound -0.0003; proxy regret -0.1713	proxy proposal selection; no identified client effect
ESConv V3	strongest frozen component fallback	1,052 conversations	utility $+0.0051$ $[0.0017, 0.0087]$; harm -0.0003 $[-0.0006, -0.0001]$	development evidence
KokoroChat V4	action + checkpoint predictor	1,671 calibration / 978 confirmation sessions	confirmation utility Brier $0.2458 \rightarrow 0.2398$; difference -0.0059 $[-0.0093, -0.0024]$	predict; recommend/execute withheld
CounselBench selection	heterogeneous judge panel	100 questions / 400 candidates	full vs panel utility $+0.0022$ $[-0.0073, 0.0126]$; V3 returns 100/100 exact fallback	higher authority withheld
MHDialog	heterogeneous judge panel	148 independent dialogues	148/148 exact fallback; utility and harm differences exactly zero	independent fallback confirmation
CaSiNo	validation-selected outcome-greedy policy	206 held-out dialogues	utility $+0.0008$ $[-0.0099, 0.0129]$; harm $+0.0269$ $[-0.0113, 0.0691]$	recommend withheld
PersuasionForGood	validation-selected effect graph	204 independent dialogues	utility $+0.0051$ $[-0.0637, 0.0748]$; harm $+0.0225$ $[-0.0061, 0.0503]$	recommend withheld
DialOp	validation-selected binary authority	83 held-out dialogues / 3 domains	utility -0.0010 $[-0.0083, 0.0061]$; harm -0.0031 $[-0.0214, 0.0100]$	recommend withheld under shift
ToolSandbox	same bare model and scenario	80 scenarios $\times$ 4 configurations; 105 confirmation	frozen-80 $+0.0391$ $[0.0239, 0.0531]$; confirmation $+0.0380$ $[0.0065, 0.0684]$	execute
ToolSandbox blind causal induction	bare, opaque lexical, random probe, source oracle	18 discovery + 18 disjoint prediction + 120 constructed tool contexts	6/6 registered edges; call success $0.400 \rightarrow 1.000$; utility $+0.587$ under fixed charges	registered precondition $\rightarrow$ repaired call; prospective component
ToolSandbox transition memory	zero-guard schema; exact learned protocol	20 discovery + 20 prediction-gate + 40 action-gate scenarios; 3 outer cells	four factors promoted; action-gate turns -5.20 $[-6.20, -3.90]$; outer gains $+0.0216$ and $+0.0759$ with positive CIs	discovered factor $\rightarrow$ execute; retrospective
τ3	same bare agent	80 paired tasks	pass^1 $0.8500 \rightarrow 0.6875$; write violations fall to zero	mandatory policy rejected; safety–utility trade-off retained

OBSERVE/PREDICT/RECOMMEND/EXECUTE promotion, fallback-relative utility–harm evidence, and externally witnessed commit.

D.2 COMPLETE SPECIALIZED-DOMAIN EVIDENCE LEDGER

Table 7 reports every domain used to motivate or test the architecture. Positive, exact-fallback, and failed gates are retained. The "status" column is the maximum authority supported by that experiment, not a post-hoc quality label. Sources: TherapyBench pairs are the RAVR-S lineage set (Pelekhov, 2026); AnnoMI (Wu et al., 2022); MentalAlign (Badawi et al., 2026); CounselBench (Li et al., 2026); EPITOME (Sharma et al., 2020); Graph2Counsel (Mandal et al., 2026); ESConv (Liu et al., 2021); KokoroChat (Qi et al., 2025); MHDialog, released with MHDash (Zhang et al., 2026); CaSiNo (Chawla et al., 2021); PersuasionForGood (Wang et al., 2019); DialOp (Lin et al., 2024); ToolSandbox (Lu et al., 2025); and the τ-bench family (Yao et al., 2024; Barres et al., 2025).

The six calibration rows summarized in the main text comprise AnnoMI judge trust, MentalAlign, CounselBench adaptation, EPITOME, AnnoMI transitions, and ESConv early-outcome prediction. They contain 296,796 scored instances across five counseling-oriented datasets; all six matched Brier

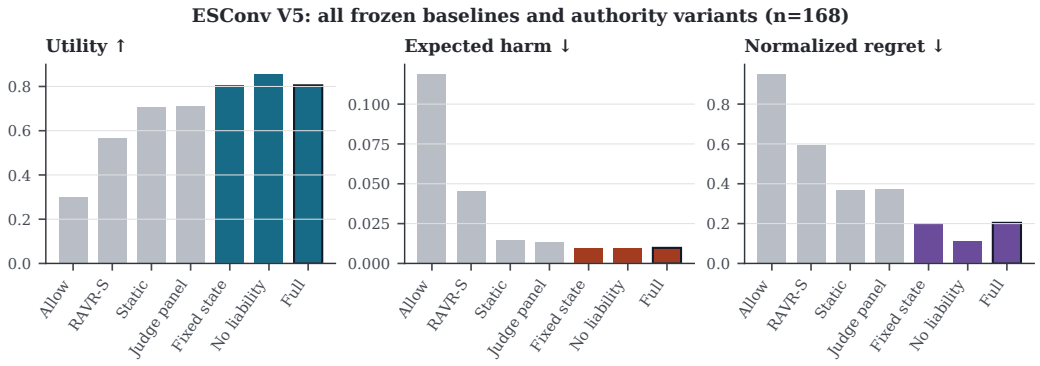


Figure 6: ESConv V5 comparison against every frozen baseline and architecture variant on the same 168 cases. The primary comparator is the heterogeneous judge panel. Fixed-state and full RAVR-W coincide because the new factor did not pass its independent action-authority gate.

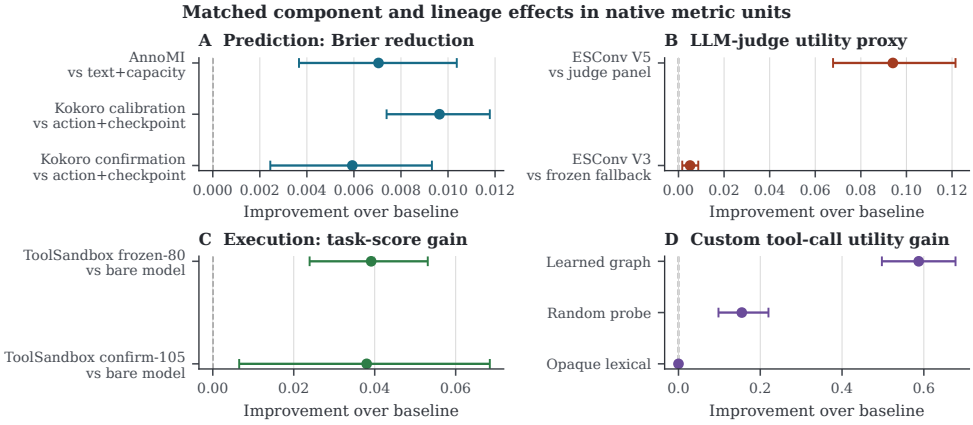


Figure 7: Grouped paired score deltas in native metrics (right is better) for prediction, proposal selection, transactional execution, and blind induction. Metrics are not pooled.

intervals favor the full component, with median relative reduction 14.7%. They are deliberately not pooled because their targets and dependence structures differ.

E Controlled mechanism audit

The deterministic audit samples 5,000 scenarios with latent utility, harm, structural validity, factor authority, uncertainty, source dependence, certificate status, overlap, and group count. It is a mechanism stress test, not an external benchmark. Table 8 includes counts and Wilson intervals omitted from the main paper.

Table 8: Controlled selection diagnostics. Intervals are Wilson 95% intervals over selected cases.

Arm	Selected	Qualified	Harmful	Invalid
Proposal only	3024	548	1683	1401
Static constraints	2704	548	1514	1081
Effect gate only	191	96	6	93
RAVR-W	36	35	1	0

RAVR-W's qualified precision is 0.972 with Wilson interval $[0.858, 0.995]$. Its safe-among-selected rate has the same interval, and its structural validity is 1.000. Qualified recall is 0.055 and switch coverage is 0.0072. These values demonstrate selectivity, not task superiority.

Nine metamorphic invariants each pass 5,000/5,000 trials: provenance alias invariance, authority-removal non-expansion, certificate-revocation non-expansion, fallback-relative authority, hard-override preservation, finite no-progress querying, rejection of offline failures as query-repairable, witness closure, and evidence-strength monotonicity. A scalar VOI gate would query in 0.0626 of scenarios, whereas strong closure queries in 0.0008; 0.0618 of scenarios are rejected because the complete frontier is not runtime closable.

F TRANSACTIONAL TOOLSANDBOX PROTOCOL

The primary ToolSandbox evaluation freezes 80 scenarios and four small-model / provider configurations. The baseline and transactional arms receive identical models, prompts, user simulators, temperatures, tools, and scenarios. The controller sees only typed tool calls and independently returned tool state. It implements eight source-level guards as `preflight` $\rightarrow$ `repair` $\rightarrow$ `commit` $\rightarrow$ `witness`; repair calls are hidden from the model, and a proposed mutation is accepted only after read-back. Task-level tool errors remain scored. Empty provider responses are classified as infrastructure failures and retried under the frozen protocol.

The independent confirmation pool was selected before outcomes by excluding the 80 primary scenarios and seven scenarios used during debugging. Its manifest SHA-256 is `881e2e749e0dfce5fa69a06b5775f9ef7d5d64a3f582a84b247e0c6408adae57`; all 105 baseline and 105 transactional episodes are retained. One empty provider response was retried, while wrong-tool and wrong-argument actions remain task failures. The paired analysis resamples scenarios, reports win/tie/loss counts and turn and latency deltas, and separately audits minefield violations.

The pre-expert freeze receipt is content-addressed and includes the four-model analysis, independent confirmation, full component ablations, mechanism receipts, all 636 passing tests, and frozen method sources. The external gate requires a positive scenario-clustered interval, positive mean effects in all four configurations, no minefield increase, and an independently positive confirmation interval. Three of four configuration-specific intervals exclude zero; the Baidu interval is borderline, so it is reported without a per-arm significance claim.

F.1 PRE-REGISTERED FOUR-ARM SETTINGS CAMPAIGN

The campaign freezes, before any outcome, all 60 ToolSandbox scenarios whose tool allow-list lies in the settings domain (ten base families, each with distraction-tool and four scrambling variants), four arms, one model (DeepSeek V4 Flash for agent and user, temperature 0), a 10-message cap, the primary endpoint (full kernel minus bare), and SHA-256 hashes of every method source; the analysis refuses to run if any hash changes. Intervals resample the ten scenario families. No scenario or run is excluded, and no arm raised an exception.

Table 9: Four-arm settings campaign (60 scenarios). Δ is relative to the bare agent; intervals are family-clustered bootstrap 95% intervals.

Arm	Similarity	Δ vs. bare [95% CI]	W/T/L	Commits
Bare agent	0.711	—	—	0
RAVR-W transactional executor	**0.865**	$+\mathbf{0.154}$ $[0.050, 0.262]$	32/19/9	—
Static intent + learned repair	0.617	-0.094	—	0
RAVR-W live full kernel	0.607	-0.104 $[-0.330, 0.110]$	22/10/28	23
Exploratory stratum: explicit low-battery enable tasks ($n = 18$)				
Full kernel	0.938	$+0.214$ (scenario CI $[0.115, 0.323]$)	15/0/3	—
Transactional executor	0.943	$+0.219$ $[0.005, 0.350]$	15/0/3	—

The full kernel's losses concentrate in `cellular_off` and `wifi_off` (0.24 versus 0.90 for bare) and in implicit requests, where its typed goal binder finds no authorized goal and fails

closed. Frozen and exploratory analyses are `analyze_toolsandbox_kernel_v3.py` and `analyze_toolsandbox_kernel_v3_scope.py`.

Fallback-exact rerun. The v3 boundary blocked every call it could not warrant, contradicting the resolver definition in §4.2. A second frozen campaign (v4) restores exact fallback—unwarranted, out-of-scope, scrambled, or multi-call proposals execute as proposed—with the kernel, protocol, learner, and prompts unchanged, and reruns the bare, static, and full-kernel arms on the same 60 scenarios. The defect was identified from v3, so v4 is development-informed; both freezes and results are released.

Table 10: Fallback-exact campaign (v4), 60 scenarios. The primary interval is family-clustered; the scenario-level interval was a pre-registered secondary analysis; the sign test is exploratory.

Comparison	Δ	Family CI	Scenario CI	W/T/L (sign p)
Full kernel − bare	$+\mathbf{0.058}$	$[-0.017, 0.151]$	$[\mathbf{0.012}, \mathbf{0.106}]$	22/30/8 (0.016)
Full kernel − static	$+0.011$	$[-0.006, 0.029]$	$[-0.004, 0.031]$	6/46/8 (0.79)
Static − bare	$+0.047$	—	$[-0.004, 0.098]$	19/30/11 (0.20)

Held-out replication (v5). The unchanged v4 method was then frozen and run against a fresh bare agent on all 184 ToolSandbox scenarios from base families absent from v3/v4 whose tool list contains a settings setter and whose name encodes an initial precondition state (23 families). Full kernel minus bare is $+0.001$ (family-clustered $[-0.016, 0.018]$; scenario-level $[-0.018, 0.019]$; 14/156/14; sign $p = 1.0$). The kernel issued one warranted switch: these tasks request a downstream action (weather, messaging, reminders) rather than the enabling setting, which the current goal binder does not bind, so almost every call follows exact fallback. Two shards that hung on a provider request were rerun in full before any outcome was computed; the retry log is released.

Precondition closure (v6). After v5 showed that the kernel almost never engaged on downstream tasks, v6 extended its typed scope to every action heading a promoted BCTI factor (`get_current_location`, `search_holiday`, `send_message_with_phone_number`) and to model-proposed enabling calls. The proposed call is the authorized intent; only learned repairs, each followed by a read, may precede it; exact fallback applies otherwise. The learned protocol, kernel, and prompts are unchanged. v6 was frozen and run against a fresh bare agent on the same 184 scenarios, so it is development-informed; 12 shards per arm were increased to 24 before any outcome existed.

Table 11: Live full kernel with precondition closure (v6), 184 scenarios in 23 families.

Quantity	Value
Bare / full-kernel mean similarity	0.464 / $\mathbf{0.575}$
Δ (family-clustered 95% CI)	$+\mathbf{0.111}$ $[0.052, 0.177]$
Scenario-level 95% CI	$[0.086, 0.137]$
Wins / ties / losses (sign test)	79 / 83 / 22 ($p = 1.0 \times 10^{-8}$)
Families with positive mean Δ	11 / 23
Warranted switches / witnessed commits	117 / 117
Exceptions (bare / kernel)	0 / 0

F.2 Witness-induced transition memory

The open-state analysis removes all executable guards from the initial schema. For every baseline trajectory, it searches only for the ordered pattern

$$\textsc{Fail}(a),\ \textsc{Read}(f = v^-),\ \textsc{Repair}(f = v^+),\ \textsc{Success}(a).$$

String-valued payload arguments are excluded from factor scope, preventing a phone number or message body from becoming a spurious precondition. The 80 Relace scenarios are partitioned by interface perturbation: 20 ordinary or three-distraction variants for discovery, 20 ten-distraction or

all-tool variants for predictive gating, and 40 argument/tool scrambling variants for the action gate. No scenario contributes to two tiers. The content-addressed manifest identifies the discovery source as bare-agent baseline conversations without the RAVR-W controller and counts support by unique scenario. Source guards are opened only after held-out recurrence, then used as a semantic validation gate. Tool errors explicitly name the blocking state, so this experiment tests authority transport rather than difficult open-ended semantic induction.

F.3 LIVE GAIA2 OPEN-STATE AUTHORITY TRANSPORT

We replay one official Gaia2 (Froger et al., 2025) adaptability scenario, `scenario_universe_21_5e0gvz`, in a local deterministic environment snapshot. Unlike the retrospective ToolSandbox analysis, a live frozen model proposes every observation and effect. The initial state schema contains no identity for the requested "Film Producer in Stockholm." RAVR-W enumerates the bounded contact store, grounds a unique typed person binding, and uses only observed fields to close calendar-attendee and email-recipient arguments.

After the first commit, two inbox notifications arrive. The pending conditional obligation prevents task closure until the inbox is observed. A reply from the grounded contact introduces a new date, revokes the predecessor event, widens the calendar read to the full proposed day, and defines a finite effect domain for "cancel them all." Every member of that witnessed domain must be deleted before the replacement event can commit; an additional email is withheld because the original send authority has already been consumed.

Table 12: Live full-chain diagnostic on one Gaia2 scenario. Intervals are Wilson 95% intervals over three stochastic runs and describe run variability only.

Condition	Successes	Success rate	Wilson 95% CI
Default ARE agent	0/3	0.000	[0.000, 0.561]
RAVR-W	2/3	0.667	[0.208, 0.939]

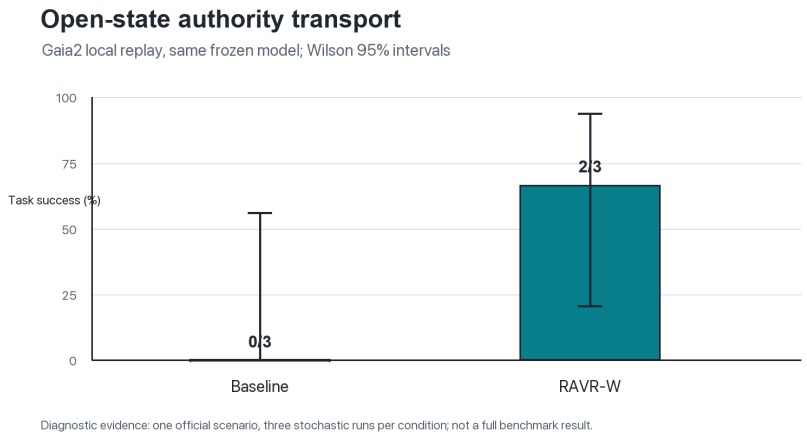


Figure 8: Matched-model Gaia2 local replay. This is an existence test of the complete open-state-to-execution path, not a full benchmark result.

The final comparison uses the same Gemini 2.5 Flash Lite agent, GPT-4o-mini judge, shared parser, tools, and scenario state. Machine-readable run rows, traces, hashes, and the analysis script accompany the artifact. The scenario was used while developing the mechanism, the snapshot was reconstructed from the official trace, and the email renderer includes an evaluator-compatibility canonicalizer. We therefore do not use this result to claim Gaia2 generalization or leaderboard performance.

Two further official development diagnostics exercise distinct typed paths. For `scenario_universe_25_cbckw0`, a relative time window identifies the target event,

participant, Messages identity, and cab date. A later cab cancellation opens a second obligation epoch. Under matched Qwen3-30B-A3B runs, the default agent scores 0/3 and RAVR-W scores 3/3 with no exception or missing validation. For `scenario_universe_23_ekcbfh`, two successive external calendar mutations introduce previously absent event creators. V22 keeps delta observation open until each event is witnessed, transports creator provenance through an exact contact conversion, spends one notification budget per creator, revokes overlapping authority before replacement, constructs only the protected-interval complement, and preserves unaffected residuals. Its final development trace scores 1.0 with no tool exception.

These are development-informed existence results. In the strict 12-scenario V10 development comparison, default scores 0/12 and RAVR-W scores 1/12, with nine semantic failures and two timeout/runtime failures. Later repair of the event-source scenario is not substituted into that frozen denominator. The complete failure matrix, traces, and SHA-256 manifests accompany the artifact.

Figure 9 records the later frozen development-panel progression under one shared small-Qwen model, judge, timeout, and observation contract. It diagnoses closure mechanisms; it is not an untouched estimate.

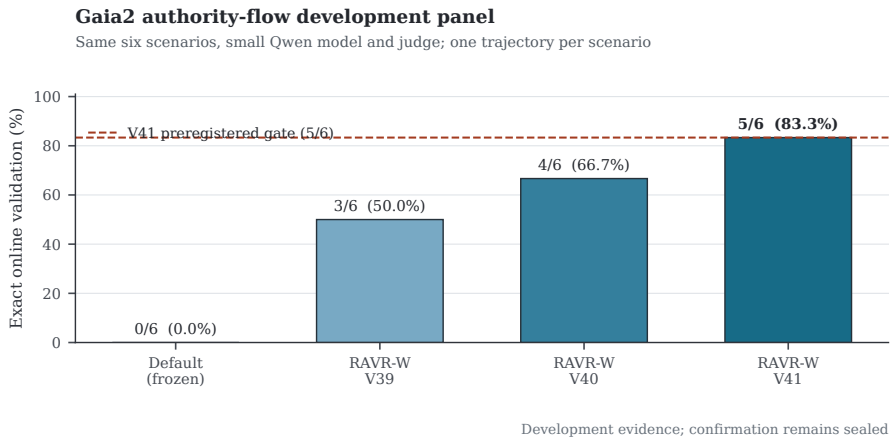


Figure 9: Frozen iterations on the same six-scenario Gaia2 adaptability development panel. Each bar contains one trajectory per scenario. The panel informed development, and confirmation remains sealed.

The sixth panel scenario was not silently discarded. V44 timed out because plural catalog matching left the initial order ambiguous; V45 fixed that binding but encoded an absent optional attachment as an empty path; V46 fixed nullable arguments but routed physical `MessagingAppV2` notifications to the wrong logical channel. V47 used receipt provenance to observe both consultant replies and completed every structured effect: the exact two-item order, two separate consultation messages, a uniquely bound red backpack, exactly one predecessor cancellation, an exact three-item successor totaling $67.97, and an email to the contact whose observed job is Data Analyst. A judge-only V48 replication produced the same structured effects.

Both official scores remain 0. The Qwen and Gemini validators rejected "The total for my order is $67.97. Could you please analyze my spending?" against "The total from the order is $67.97. Please analyze my spending." Recipient, subject, amount, analysis request, and signer match. V47 took 21.1 seconds; V48 took 635.1 seconds, a 30.1-fold ratio, because the nominal 420-second scenario timeout did not bound soft validation. This is evidence of one cross-judge false negative and an unbounded evaluation stage, not an error rate estimate. We retain the official failures and use them to motivate the typed execution-lineage certificate; no oracle phrase is compiled into the agent.

Prospectively frozen V49 changes only this authorization boundary. Before the email effect, the runtime emits one trace-visible certificate binding the observed Data Analyst contact, the successor order receipt, the original spending-analysis intent, and the observed current-user identity. Its four argument checks pass, and a separate deterministic auditor reconstructs the certificate payload from the trace and reproduces certificate ID `effect-certificate:4a151151c00d...f2196`. The

receipt precedes the email event; the initial two-item order totals $39.98, the predecessor is cancelled exactly once, and the witnessed three-item successor totals $67.97. The primary preregistered mechanism gate therefore passes. The official score remains 0 under the same email paraphrase mismatch. Average benchmark runtime is 14.95 seconds (18.86 seconds for the job). This is one prospective development mechanism result, not a sealed transfer estimate or a relabelled official pass.

V50 then freezes one architecture over all six panel scenarios. It again returns 5/6 and preserves every V41 official pass; the Shopping trace also repeats all structured and certificate checks. V50 nevertheless fails its preregistered panel protocol because the official output records a 420-second timeout. Trace inspection exposes a general terminal-control defect: the word "which" in a declarative completion message is misclassified as a clarification request, reopening a closed person binding until the run reaches 80 model calls. Prospectively frozen V51 changes only this speech-act boundary. Its terminal audit passes all gates: no timeout or exception, one certified email, exactly one later acknowledgement, no post-email binding receipt, and 3 model calls, a 96.25% reduction from V50. The structured mechanism again passes while the official score remains 0 under the unchanged paraphrase mismatch. Thus V51 establishes terminal closure on the same repeatedly used development scenario, not transfer or a relabelled official pass.

V52 attempted the corresponding six-scenario replication but selected an exhausted provider key. Five workers received HTTP 402 responses and reached the frozen 420-second limit; the one exported Shopping trace is retained, and V52 is recorded only as an infrastructure failure. Before any recovery call, V53 froze the same model, method, code hashes, data manifest, timeout, gates, and auditors; only a positive-balance key selected by a declared rule changed. V53 completes all six scenarios in 75.7 seconds with no provider, timeout, runtime, or tool exception. The single architecture returns 5/6 exact official validations versus 0/6 for the frozen matched default, preserves all five prior passes, and passes both the structured Shopping mechanism and terminal audit. The Shopping trace contains one recomputable certificate, one email, exactly one later acknowledgement, no post-email binding receipt, and 3 main-agent calls. Its official score remains 0 under the same literal paraphrase mismatch. Thus V53 passes the complete prospectively registered development-panel gate; it remains development evidence rather than an untouched transfer estimate.

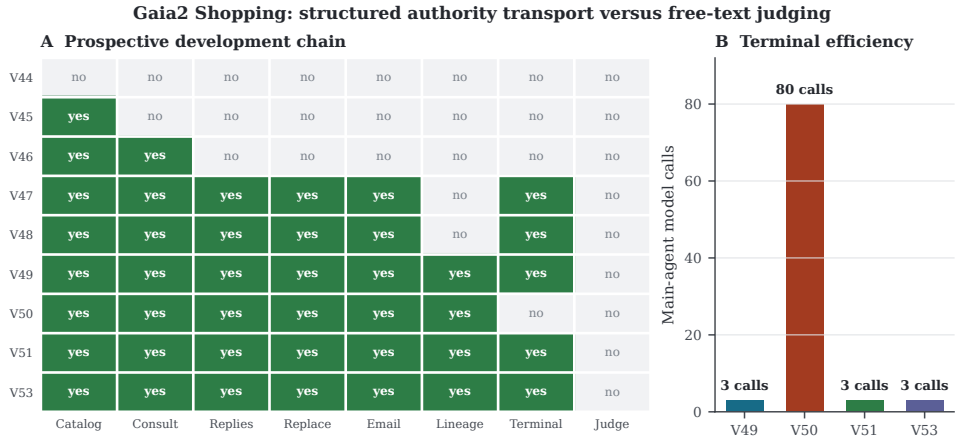


Figure 10: Prospectively frozen Shopping development revisions. Green cells are structured effects established by trace audit; the final column is the official free-text judge. V49 adds a recomputable lineage receipt. V50 repeats the mechanism but fails terminal closure; V51 prospectively closes the loop. V53 repeats lineage and closure inside the passing six-scenario panel with 3 rather than V50's 80 model calls. Official Shopping failures remain unchanged.

On the 40 action-gate scenarios, learned preflight and repair change official similarity by $+0.0129$ $[-0.0377, 0.0465]$, above the fixed -0.05 non-inferiority margin; turns change by -5.20 $[-6.20, -3.90]$ and minefield similarity by exactly zero. All four factors activate. Ten polarity-inverted or cross-state controls receive zero promotions.

Table 13: Full open-state-to-execution receipt. Supports count independent failure–repair–retry traces. The action gate is shared because execution authority is granted to the compiled four-factor protocol.

Learned factor	Discovery	Predictive gate
cellular disabled ⊣ send message	7	8
low battery ⊣ enable cellular	4	4
low battery ⊣ enable location	3	4
low battery ⊣ enable wifi	2	3

The outer audit uses no Relace outcomes for model/provider selection. In each of the StreamLake, Baidu, and Qwen/Alibaba cells, receipt inspection finds only the four learned source-guard identities; the other four source guards never activate. Consequently the compiled learned protocol is exactly equivalent on all 240 outer trajectories, not merely similar in aggregate. Paired official similarity changes are respectively $+0.0216$ $[0.0143, 0.0293]$, $+0.0304$ $[-0.0005, 0.0609]$, and $+0.0759$ $[0.0224, 0.1250]$. Because the executions predate this analysis, this is retrospective external transport rather than a prospective confirmation.

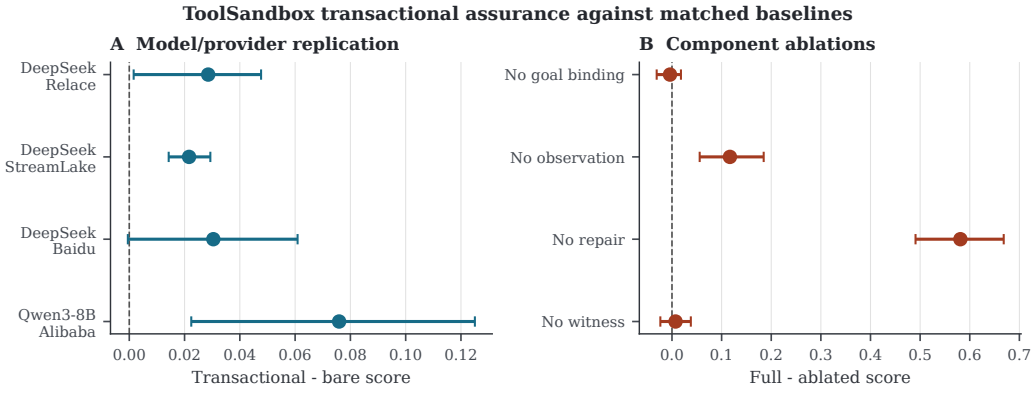


Figure 11: ToolSandbox replication and component evidence. Left: transactional minus bare-agent score across four frozen small-model/provider configurations. Right: full controller minus each ablation. Observation and repair drive task score; goal binding and commit witness primarily encode assurance on this set.

F.4 Prospective blind causal transition induction

The confirmatory induction study was frozen before its external executions. The public manifest commits to both method-source hashes, the SHA-256 of a separate outcome key, all 156 initial states, randomized candidate-channel orders, partitions, baselines, and analysis endpoints. The immutable freeze hash is `df99e21100311c0f7eba7ef51c85b2c1d70754c3ae7c7e9a4afa967c47b64125`. Every probe creates a fresh ToolSandbox `ExecutionContext`; the learner observes the initial state channel, intervention success, and final action success, but receives neither exception text nor exception class.

Six action families are tested: enabling cellular, Wi-Fi, or location under low-battery constraints; messaging under cellular constraints; location lookup under location-service constraints; and holiday search under Wi-Fi constraints. The authors construct the states and register four Boolean channels, getters, and channel-specific repair calls before probing. The learner selects edges within this registry; it does not name previously unseen semantic variables. Each factor requires three independent discovery witnesses and three disjoint prediction witnesses, exceeding the frozen minimum of two in each partition. All six source preconditions are recovered with precision and recall 1.000 within the registered candidate set. The 120 action contexts contain 72 blocked and 48 executable initial states; compound blockers are reserved for this action gate rather than used for factor selection.

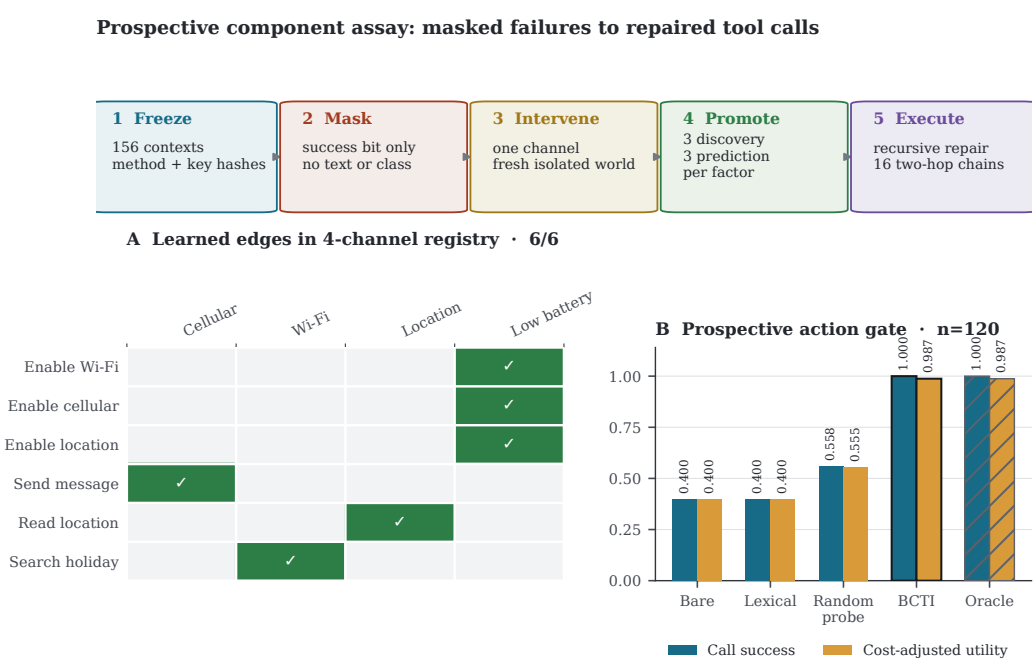


Figure 12: Prospective precondition-chain component evidence. Top: the hash-frozen protocol masks failure payloads, identifies preconditions by isolated interventions, requires disjoint recurrence, and executes only the promoted recursive graph. Bottom left: all six source preconditions are recovered without a false edge. Bottom right: BCTI matches source-oracle call success and utility on 120 held-out, author-constructed contexts. The $+0.587$ utility gain over bare execution is descriptive for this designed context mix.

Table 14: Prospective masked-failure tool-call assay, not official task success. Utility uses prespecified charges for reads, interventions, and failed retries. Intervals are paired episode bootstraps conditional on six fixed action families and the designed blocked/unblocked mix, not estimates of transfer to new tasks.

Arm	Success	Utility	Utility gain [95% CI]
Bare action	0.400	0.400	—
Opaque lexical	0.400	0.400	0.000 [0.000, 0.000]
Single random probe	0.558	0.555	0.155 [0.098, 0.220]
BCTI learned graph	**1.000**	**0.987**	**0.587 [0.497, 0.678]**
Source oracle	1.000	0.987	0.587 [0.497, 0.670]

Sixteen contexts require a genuine two-repair chain, such as disabling low battery mode before enabling cellular and only then sending a message. Every repair is followed by a state read before its dependent edge can execute. The learned protocol's content hash is `98dae0f593fa5eb20d782a9bc807e6dbe09851a4b0768db40bef9df265dddf70`. This is a component-level action assay over actual ToolSandbox tool implementations; official end-to-end model scores remain the separate matched experiments above. A call counts as successful if it raises no exception; only repair states receive read-back checks. No final semantic task-effect witness is tested. The learned and source-oracle graphs tie on all 120 cases, so this study establishes transport of registered preconditions, not superiority to correct static rules.

G KokoroChat frozen study contract

The first external study uses the Japanese originals. It excludes evaluation IDs at or below 1100 because low-ID public examples were inspected before the freeze. It also excludes records after

the frozen collection endpoint. The chronological split was fixed before row-level outcomes were opened:

Table 15: Outcome-blind KokoroChat split.

Partition	Sessions	Decision points
Development (2020–2022)	3064	9192
Calibration (2023)	1671	5013
Confirmation (2024-01-01 to 2024-09-08)	978	2934

Three decision points are fixed at 25%, 50%, and 75% of each session. The twenty client-role feedback items define the observational utility contract; the three serious-issue screens remain separate harm outcomes. The independent unit is the session. Because counselor actions were not randomized, the study tests selective prediction, transport, and observational policy value under explicit assumptions; it does not establish a causal treatment effect.

The V1 observer pilot exposed a serialization defect and produced 0/24 parseable records. A versioned amendment moved allowed-value arrays outside scalar output fields without changing factors, actions, thresholds, or outcomes. The pinned DeepSeek/DeepInfra campaign then completed 9,192/9,192 development records from 3,064 sessions, with 466 abstentions and cost USD 1.8982. It exports only typed factors, action families, uncertainty, and evidence indices. Dialogue text, translations, row-level feedback, and raw hosted-model responses are not public artifacts. The frozen V1.3 analysis produced zero action certificates because no action-scope cell passed utility, harm, and overlap jointly. V4 therefore separates four authority tiers: observe, predict, recommend, and execute. All 15 non-empty subsets of the universal components were compared only on development data. The selected subset—open-world state, information gathering, and state-dependent intervention—was frozen before chronological calibration and independent confirmation observer outputs and outcomes were opened.

On 1,671 calibration sessions, the frozen subset reduced utility Brier by 0.0096 with a session-bootstrap interval $[-0.0118, -0.0074]$ and satisfied the predeclared harm margin. On 978 confirmation sessions it reduced utility Brier from 0.2458 to 0.2398, a difference of -0.0059 $[-0.0093, -0.0024]$, while harm Brier also decreased. Thus prediction is authorized, but recommendation and execution remain withheld because predictive performance does not supply action-effect evidence.

A separately frozen 120-session observer audit completed all 360 decision points with both DeepSeek-v4-flash on DeepInfra and Qwen3-30B-A3B on SiliconFlow. Action-family agreement was 0.8722 and abstention agreement was 0.9639. Directly substituting Qwen's state coordinates did not preserve the utility gain. An exploratory, outcome-blind, five-fold coordinate transport changed the utility difference from $+0.0109$ to -0.0012, but its interval still crossed zero. This motivates observer-indexed state coordinates and a new prospective transport confirmation; it is not used to strengthen the primary claim. A registered Mistral replication remains incomplete because of provider rate limiting and is excluded from estimates.

H PROFESSIONAL VALIDATION BOUNDARY

The professional study intentionally avoids a generic preference comparison between two free-form answers. It builds a blinded semantic reference for the paper's full mechanism. Each case is a three-checkpoint trajectory. Qualified professionals independently select every permissible action and the single best next action at S0, S1, and S2; at S1 they identify the minimal evidence set, and at S2 they label whether the earlier decision should be maintained, weakened, revoked, or replaced. System outputs are generated and hash-frozen before human labels are opened, and method identities, scores, prompts, and predictions are never shown to raters.

The development pilot covers counseling support, clinical triage, and consumer finance. Each domain has 15 development trajectories and five qualified professionals. A balanced block allocation assigns nine trajectories to each professional and three independent ratings to every trajectory. A strict majority defines each field; unresolved fields receive a fourth blinded domain adjudication. The pilot

estimates timing, missingness, raw agreement, paired discordance, and the confirmatory sample size. Pilot cases and labels are excluded from confirmatory effect estimates.

The primary endpoint is trajectory-chain success: all three selected actions and the S1-to-S2 revision must match the locked expert reference. We separately report action accuracy, membership in the expert-permissible set, unsupported action rate, revision accuracy, and evidence-span F1. Systems are paired on the same trajectories; uncertainty resamples trajectories within domain, and the three domains receive equal weight. A multidomain superiority claim additionally requires no increase in unsupported actions and the same effect direction in all domains. These judgments establish semantic and professional validity of a next action under the supplied evidence; they do not establish clinical or financial outcome efficacy.

I Claim ledger and reproducibility receipt

Claims are tagged as theorem, mechanism, lineage, confirmatory, or human. Theorem language always states its assumptions. Controlled diagnostics are not reported as benchmark performance. Predecessor component results motivate the design but cannot become end-to-end RAVR-W claims. External performance requires a frozen method hash, development-independent groups, matched information and candidates, grouped paired intervals, retained infrastructure failures, utility superiority or non-inferiority, no higher harm, and useful coverage.

The current checkpoint contains 636 tests in the pinned tau3 environment. The receipt hashes the kernel, state bridge, typed authority lattice, contingent planner, paper, claim ledger, KokoroChat freeze, observer contract, mechanism audit, ToolSandbox analyses and frozen method sources, outcome-blind manifests, and full test receipt. The counts are software verification evidence, not empirical performance.

J Identification boundary and grouped precondition probes

No-free-effect-authority corollary. Let $A \in \{a, b\}$, $Y \in [0, 1]$, $p_j = P(A = j)$, and $\mu_j = E[Y \mid A = j]$. Assume consistency but no conditional exchangeability or restriction on missing potential outcomes. By the worst-case bounds of Manski (1990),

$$L = p_a\mu_a - p_b\mu_b - (1 - p_b) = p_a(\mu_a - 1) - p_b\mu_b \le 0, \qquad U = p_a\mu_a + (1 - p_a) - p_b\mu_b.$$

Both endpoints are attained by setting the unobserved potential outcomes to their extremes. Thus an observation-only gate cannot certify $E[Y(a) - Y(b)] > 0$ uniformly over compatible structural models. More precisely, if a randomized gate's probability of a false strict-improvement certificate is at most α for every compatible model, then its certification probability on any one observational equivalence class is at most α: that class contains a model with nonpositive effect, and all models in it induce the same gate input. The conclusion does not apply after randomization, justified identifying assumptions, or direct intervention evidence, and does not prohibit rule- or consent-authorized actions.

An exact audit uses $p_a = p_b = 1/2$, $\mu_a = 3/4$, and $\mu_b = 1/4$. Every model in a one-parameter family has the same observed joint law and oracle observational probabilities. Assigning missing outcomes so that $P(Y(a) = 1 \mid A = b) = t$ and $P(Y(b) = 1 \mid A = a) = 1 - t$ yields $E[Y(a) - Y(b)] = t - 1/4$, from $-1/4$ to $3/4$. Figure 13 shows why a larger observed reward for a is not an effect certificate. This is an exact toy mechanism audit, not evidence about counseling effects.

Rollback does not identify individual welfare. Let a tool action set a monitored bit to one and send an external notification; rollback restores the bit to zero. The monitored state then equals the fallback state, yet the notification remains visible. A transient harm gives the same counterexample even without communication. Accordingly, Proposition 8 is a declared-witness commit property only. It cannot be read as deterministic person-level non-inferiority. Isolated pre-commit trials are permitted only under a complete side-effect registry and a genuinely resettable, external-effect-free boundary; live irreversible actions require prospective permission and harm evidence. Reset policies in RL already study the difficulty of reversibility, so we do not claim reset itself is novel.

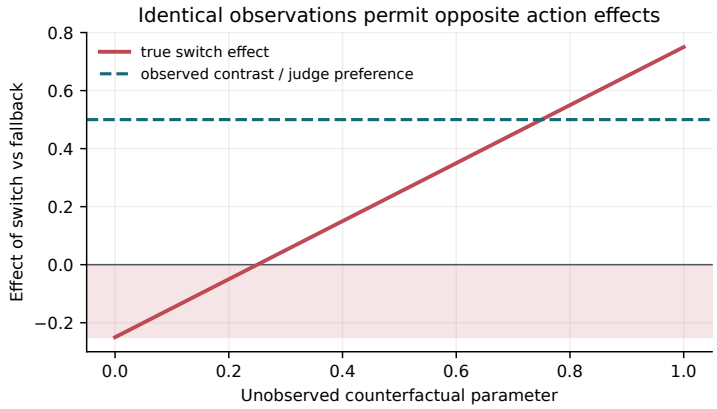


Figure 13: Identical observational data and judge probabilities permit opposite effects of switching from a fallback. The shaded region is true expected harm from the switch.

Compound blockers as a group-testing oracle. Suppose a registered action succeeds iff all channels in an unknown set $R \subseteq [m]$, $|R| \leq k$, are at known good values. In a fresh isolated world, repair every candidate except a disabled set G and retry the action. The success bit is one iff $R \cap G = \varnothing$. This is the noiseless adaptive group-testing query after outcome complementation (Aldridge et al., 2019). Our grouped splitter uses a chunk of roughly n/d among n remaining candidates and d remaining possible blockers, then binary-searches a failing chunk. It proposes factors only; disjoint validation and the ordinary warrant gate still control promotion and use. The current implementation has no claim to Hwang's exact optimal test bound. The old singleton assay cannot detect $k > 1$ simultaneous blockers from an all-bad initial state because no single repair yields success. A passive success-intersection comparator is included but is not the SAM learner of Stern & Juba (2017).

In 100 seeded noiseless synthetic worlds per (m, k) cell, with $m \in \{4, 8, 16, 32, 64, 128, 256\}$ and $k \in \{1, 2, 3, 4\}$, grouped probes recovered the exact set in all 2,800 cases. At $m = 256$ they used mean 13.0, 21.4, 29.1, and 36.1 action attempts for $k = 1, 2, 3, 4$; matched-budget passive exact recovery was 0.12, 0.10, 0.01, and 0.00. The old singleton assay recovered only $k = 1$ there and used 257 probes. A stronger matched leave-one-out baseline repairs all channels, then omits one per fresh reset; it recovers all 2,800 compound and singleton sets. At $m = 256, k = 4$ it needs 257 action attempts versus 36.11 grouped, while at $m = 4, k = 4$ it needs 5 versus 10. Thus the advantage concerns probe count in large sparse registries, not universal active-learning accuracy. This controlled mechanism test assumes known channels and repair values, deterministic conjunction, noiseless success, and exact reset; it does not establish external benchmark or human-outcome gains.

For a separate noisy-bit stress test, the number of independent odd repeated queries is the smallest r for which the exact binomial majority-error tail times the conservative logical-query cap $m + k(2 + \lceil \log_2 m \rceil) + 3$ is at most 0.05. The extra k queries falsify each proposed necessity claim by disabling that channel alone after all other repairs; a simple OR precondition is rejected. Across 100 generated worlds for each of 12 cells with $m \in \{16, 64, 256\}$, $k \in \{2, 4\}$, and symmetric independent bit-flip rates $\epsilon \in \{0.05, 0.10\}$, exact recovery including fail-closed cases ranged from 96/100 to 100/100. At $m = 256, k = 4, \epsilon = 0.10$, the derived $r = 13$ required 471.0 mean raw attempts. This is a known-noise synthetic test, not evidence that real tool errors are symmetric or independent. Three additional development workflows composed actual ToolSandbox functions with two simultaneous blockers each; all three source-known blocker pairs were recovered in 7–8 attempts per discovery or prediction partition. Recurring factors were passed to the existing executor; all three actions then succeeded after two repairs. These source-inspected workflows are not unopened benchmark tasks and do not check communication witnesses or individual harm.

Mutation cost and clone assumption. Action attempts are not total work. At $m = 256, k = 4$, rebuilding each fresh all-bad world requires 8,305.08 mean channel interventions for 36.11 action attempts. If the environment can supply isolated clones of a verified all-good target state, disabling

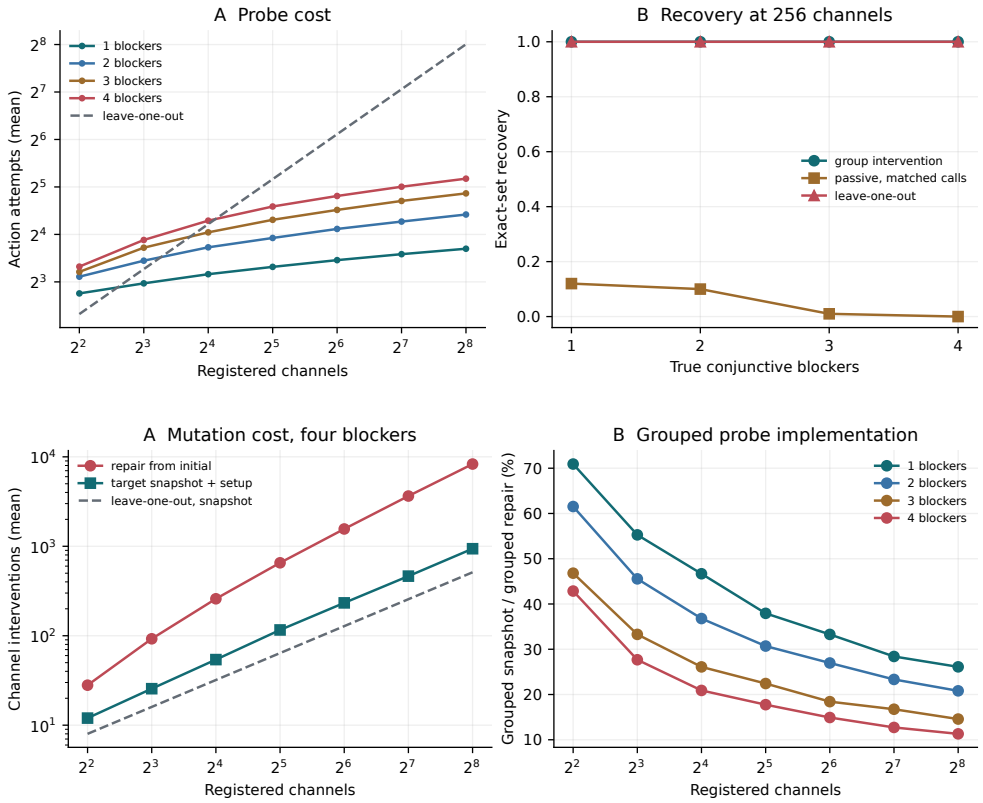


Figure 14: Synthetic conjunctive precondition assay. Top: action-attempt cost and exact recovery; both active methods recover all blocker sets, but grouped probing uses fewer attempts only in sufficiently large sparse registries. Bottom: channel-intervention cost, with one-time target-snapshot preparation and free isolated cloning assumed. Leave-one-out is cheaper in snapshot mutations at $m = 256, k = 4$ even though it needs more action attempts.

only the tested set has a modeled cost of 939.08 interventions including one-time preparation of all 256 channels and assuming free isolated cloning, while recovering the same factors with the same number of action attempts. Matched leave-one-out uses 65,536 interventions from all-bad worlds, or 512 including setup under the same free-clone snapshot assumption. Grouping therefore saves action attempts but uses more snapshot mutations. The snapshot mode is unavailable for live worlds with uncontained external effects, and clone construction itself may have costs not modeled here. Figure 14 exposes the probe/mutation trade-off.