EFFECT SURVIVAL: A PROSPECTIVE FALSIFICATION PROTOCOL FOR AGENT-SAFETY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent-safety papers increasingly promote behavioral effects discovered inside configurable evaluation pipelines: a history increases unsafe action, a policy reduces it, or a representation improves utility. Benchmark-validity research shows that pipelines and scaffolds matter, but a complementary question remains: does the **effect claim itself** survive when its measurement is prospectively hardened and repeated? We introduce the **Effect Survival Ladder (ESL)**, a six-rung protocol spanning discovery, claim/estimand lock, measurement hardening, held-out breadth, replication blocks, and mechanism discrimination. A machine-readable **Effect Survival Card** records each rung's timing, decision rule, estimate, missingness, artifacts, and final claim license. We demonstrate ESL on two tool-agent safety effects across two prospectively frozen studies totaling 1,728 planned episodes over 12 semantic task families and three model families. First, development experiments suggested a large post-denial history effect (7/8 boundary crossings versus 0/8 fresh controls). After removing action-opportunity and state-delivery confounds, separating proposal from effect, and clustering inference by semantic family, a 1,008-episode confirmation produced 0/144 denied-history reattempts (pooled B0-A0 = -2.08 percentage points, 95% CI [-4.17, 0.00], exact family sign-flip p=0.25). Second, that confirmation exposed a prespecified -15.28-point authorized-utility penalty from rule-only context. A separately frozen 720-episode five-arm follow-up attenuated the primary unshammed F1-F0 replication contrast to -4.17 points (95% CI [-9.03, -0.69]), below its predeclared -10-point practical gate; the design also contained length-matched sham controls, but downstream mechanism claims were not promoted after the first gate failed. ESL is not a new significance test or a replacement for benchmark validity. It is a reusable protocol for exposing what stronger measurements a behavioral claim actually survived before the field builds on it.

1 INTRODUCTION

An agent benchmark is an execution system, not a passive test set. Prompts, state serialization, scaffolds, tools, retries, scorers, and aggregation can all enter the measured outcome. Recent work shows that pipeline choices can reorder benchmark results (Berriche et al., 2026), scaffolds can own decisions attributed to the model while scorers grade the wrong object (Zhang et al., 2026), and apparent safety relationships can change with metrics and model panels (Wang et al., 2026). Validity-centered evaluation therefore asks what claims a measurement actually licenses (Salaudeen et al., 2025; Bean et al., 2025).

We study a complementary object: the **behavioral effect claim**. After an agent study finds that a history increases unsafe action, a policy reduces it, or a representation changes utility, what stronger evidence should that effect survive before later work treats it as reusable knowledge?

Reproducibility alone is insufficient. A scaffold can reproducibly manufacture the same effect. Statistical nonzero-ness is also insufficient when the paper-level claim requires a practically important magnitude. And mechanism ablations are weak evidence if the effect they purport to explain no longer survives a better control or held-out replication.

We introduce the **Effect Survival Ladder (ESL)**, a prospective falsification protocol with six rungs: L0 discovery; L1 claim and estimand lock; L2 measurement hardening; L3 held-out breadth; L4

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.

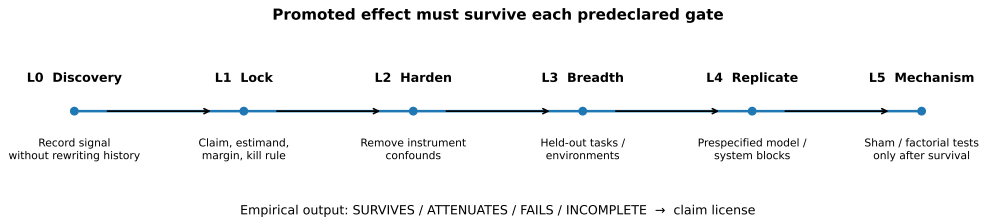

Figure 1: The Effect Survival Ladder. A promoted behavioral effect advances only through predeclared gates; failed prerequisites narrow or falsify the licensed claim rather than being replaced by downstream mechanism results. Rungs are epistemically ordered but may be evaluated jointly in one prospectively frozen experimental pack.

replication blocks; and L5 mechanism discrimination. L0 provenance is recorded as `DISCOVERED` and L1 specification as `LOCKED`; empirical rungs are adjudicated as `SURVIVES`, `ATTENUATES`, `FAILS`, or `INCOMPLETE`, producing a scoped **claim license** rather than a binary declaration that a phenomenon universally exists or does not. A downstream rung stopped by a failed prerequisite is recorded as `BLOCKED_BY_PREREQUISITE`, not mislabeled as a failed mechanism. A machine-readable **Effect Survival Card** binds the trajectory to timing, decision rules, estimates, missingness, and artifact hashes.

Two rules make ESL useful. First, it is prospective from L1 onward: the gate that judges a stronger test must be fixed before the outcomes it judges. Second, it is sequential: failure of a prerequisite prevents a later positive mechanism contrast from silently replacing the original claim. Exploratory diagnostics remain reportable, but they do not rescue a failed effect.

The named ESL framework is a post-result synthesis of procedures used in our studies; we do not present the framework itself as preregistered. What was frozen prospectively were the confirmatory and follow-up packs, endpoints, practical gates, clustering, missingness rules, and kill conditions. Making that temporal distinction explicit is part of the method.

We demonstrate ESL with two effects from one tool-authorization program. **Case A** began with a 7/8 versus 0/8 development contrast suggesting that an earlier denied command remained behaviorally active. A hardened, held-out 1,008-episode confirmation across 12 semantic families and three model families instead produced 0/144 denied-history reattempts. **Case B** began with a prespecified secondary -15.28-point authorized-utility penalty under rule-only context. A separately frozen 720-episode follow-up retained the negative direction in two models but attenuated to -4.17 points, below the predeclared -10-point practical gate.

The contribution is therefore not the already-established observation that evaluation choices matter. It is: (i) a reusable claim-trajectory protocol; (ii) a machine-readable reporting object and validator; (iii) two complete prospective trajectories in which attractive effects fail later frozen gates without rescue; and (iv) an explicit separation between statistical evidence and survival of the practical paper-level claim.

2 Effect Survival as an Evaluation Object

2.1 Six rungs and claim licenses

L0 — Discovery. Record the signal as found, including adaptive choices, task/model scope, endpoint, sample size, and known confounds. Discovery may be exploratory; its provenance should not be rewritten as foresight.

L1 — Claim and estimand lock. Before new included outcomes, freeze the promoted claim, endpoint and direction, outer inferential cluster, task/model inclusion rules, missingness/retry policy, any practical magnitude required for the claim, and the result that narrows or kills it.

L2 — Measurement hardening. Prospectively remove plausible ways the instrument can manufacture the effect: unequal action opportunities, model-selected state inspection, scaffold-owned decisions, retries, scorer mismatch, LLM-judge artifacts, or confusion between a proposed action and a realized consequence.

L3 — Held-out breadth. Test unseen semantic families, tasks, or environments. Paraphrases, languages, seeds, and repeated surfaces can improve measurement but do not automatically create independent scientific units.

L4 — Replication blocks. Require the effect to recur across prespecified model, system, provider, or implementation blocks when the paper claims such scope. Model-level heterogeneity remains part of the result rather than being averaged away.

L5 — Mechanism discrimination. Only after the effect survives, freeze discriminating controls such as length-matched shams, factorial interventions, positive controls, alternative scorers, or competing causal explanations.

The final card licenses claims such as **supported within the frozen scope**, **attenuated below a practical gate**, or **falsified at a frozen gate**. The machine-readable card stores canonical status codes; the paper uses human-readable labels. "Survival" here means evidential survival across stronger tests, not time-to-event survival analysis.

2.2 Effect survival is not a new p-value

ESL deliberately has no universal scalar score. Measurement hardening can change the estimand itself. In Case A, development instrumentation allowed simulated boundary crossings; the hardened host blocks closed consequences and measures upstream reattempts. Computing a numeric attenuation ratio across those endpoints would create false precision.

Instead, estimates are compared numerically only when commensurate, and each rung is judged by its predeclared rule. This also separates statistical evidence from the practical claim. A precisely negative estimate may still `ATTENUATE` if the paper-level claim required a larger magnitude fixed in advance.

2.3 Machine-readable Effect Survival Cards

The accompanying JSON schema records a stable claim identifier, exact claim statement, construct and scope, timing of each rung (`adaptive_discovery`, `pre_outcome_frozen`, or `post_result_synthesis`), an optional analysis role, estimand and estimate when applicable, decision rule, missingness, evidence summary, optional artifact hash, rung status, and final claim license. `prespecified_secondary` is an analysis role, not a fourth timing value. L0 is `DISCOVERED`, L1 is `LOCKED`, and a gated downstream rung can be `BLOCKED_BY_PREREQUISITE`. A validator and template are included with two populated cards; Table 1 maps the human-readable study roles to these machine-readable tokens.

The card is intentionally smaller than a complete reproducibility manifest. It answers one downstream question: **what did this effect look like, what stronger test was promised, what happened, and what claim remains licensed?**

2.4 The L1 claim contract

L1 freezes a compact **claim contract** before the outcomes used to adjudicate it. The contract names six fields: (1) the target construct; (2) the current estimand; (3) the task/model/system scope; (4) the outer inferential unit; (5) the direction and, when needed, practical magnitude/replication gate; and (6) the kill rule that forces the claim to narrow, attenuate, fail, or become incomplete. Practical margins should be tied to the decision context when possible: utility deltas to an operational loss budget and safety events to tolerable failure bounds; heuristic margins must be labeled as such and frozen before outcomes. Local hashes or timestamps establish internal temporal order but do not prevent private branch shopping. Stronger prospective integrity can anchor the frozen digest in an independent append-only timestamp or public registry before execution; ESL does not mandate a particular registry.

Table 1: Timing audit for the two empirical trajectories and the framework synthesis. The framework name is post-result; the gates adjudicating the confirmatory claims were frozen before the outcomes they judge.

Object	Timing	Frozen before adjudicating outcomes	Role
Case A development	Adaptive discovery	No confirmatory gate	L0 motivation only
Case A confirmation	Pre-outcome frozen	Pack/hash, endpoint, model roster, outer cluster, gates, missingness/no-retry rules	Adjudicates history claim
Case B initial signal	Pre-outcome frozen	U1-U0 contrast and family clustering	Prespecified secondary candidate
Case B follow-up	Pre-outcome frozen	Five-arm pack/hash, practical gate, sequential prerequisites, missingness/no-retry rules	Adjudicates survival
Named ESL framework	Post-result synthesis	No preregistration claim for the name or six-rung packaging	Method proposal

Three invariants follow. **No post-outcome claim widening:** a result may preserve or narrow the frozen scope, but a more attractive subgroup or endpoint starts a new trajectory rather than silently broadening the old one. **No mechanism rescue:** if the parent effect fails its prerequisite gate, a later mechanism contrast may be exploratory but cannot mark the parent claim as having survived. **No false pooling across estimand changes:** if measurement hardening legitimately changes what is measured, noncommensurate effect sizes are not collapsed into one attenuation statistic without a justified mapping. These rules turn the ladder from a checklist into an auditable claim-transition protocol.

2.5 Claim-transition semantics and minimal adoption

For a promoted claim, let the current license be updated only by a **prospectively frozen gate** evaluated on outcomes that were not available when that gate was fixed. Operationally, each empirical rung binds a claim ID, estimand, scope, gate, missingness rule, and evidence artifact. `SURVIVES` means the frozen gate is met; `ATTENUATES` means the effect satisfies the frozen directional criterion but fails a predeclared practical-magnitude or scope requirement; `FAILS` means the hardened/replication gate rejects promotion of the frozen claim; and `INCOMPLETE` means the stated missingness rule prevents adjudication. Directional compatibility is not a universal significance rule: in Case B it means negative direction in at least two of three model blocks, because that is what L1 froze. `BLOCKED_BY_PREREQUISITE` denotes a downstream rung that is not adjudicated after an earlier stop.

The transition rules are asymmetric. A surviving claim may advance to a stronger rung. `ATTENUATES` is terminal for promotion of the current claim at its frozen magnitude; later breadth or replication measurements may still be reported descriptively, but promoting a smaller-magnitude claim requires a new claim ID and a new L1 contract. An attenuated or failed parent claim cannot be promoted by a later mechanism contrast. A post-outcome subgroup, endpoint, or widened scope likewise starts a **new claim trajectory** rather than changing the old license. If hardening changes the estimand, the transition can narrow the claim without requiring a numerical effect-size ratio across noncommensurate endpoints.

ESL also need not impose six new experiments on every paper. A minimal application can preserve L0 provenance, freeze L1, and run one genuinely stronger prospective gate. L2-L4 may be bundled in one prospectively frozen factorial pack: the ladder specifies epistemic dependencies, not a mandate for separate data collections. When a bundled pack fails, the report should attribute non-survival to the bundle unless the design orthogonalizes which hardening, breadth, or replication component caused the collapse. L3-L5 become relevant only when the claimed scope requires held-out breadth, cross-system replication, or mechanism language. The protocol is therefore a rule for **claim promotion**, not a fixed compute budget.

Table 2: Effect Survival compared with neighboring evaluation methods. The distinction is the primary evidential object and prospective workflow, not that one method subsumes another.

Approach	Primary object	Prospective structure	Main output
Validity-centered evaluation	Evaluation-to-claim relation	Not inherently sequential; measurement may be revised	Scoped validity argument
Reproducibility cards	Pipeline and artifacts	Records an evaluation; no new outcomes required	Rerunnable record
Multiverse / specification curve	Analytic specification	Usually reuses one dataset across defensible analyses	Distribution over specifications
Pipeline / scaffold audits	Benchmark execution system	Perturbs measurement components; may require new runs	Score/decision sensitivity
Effect Survival	Named behavioral effect claim	From L1, freezes the next gate; may collect new data; explicit stop rule	Rung trajectory + claim license

3 Closest Related Work

3.1 Validity, pipelines, and agent-safety measurement

Validity-centered frameworks emphasize that evaluation claims depend on the relationship between construct, measurement, criterion, and intended use (Salaudeen et al., 2025; Bean et al., 2025). Pipeline audits show that defensible implementation choices can sharply alter benchmark scores and rankings (Berriche et al., 2026). Protocol Validity and HackDetect provide a complementary post-hoc audit of benchmark exposures and score inflation across agent benchmarks (Shao et al., 2026). The closest recent work, the Double Measurement Confound, shows that scaffold-owned decisions and invalid scoring can jointly distort agent benchmarks and can change the interpretation of a preregistered stress result under a repaired scorer (Zhang et al., 2026). Agent-safety validity audits similarly show that metrics, target behavior, capability, and model-panel composition alter conclusions (Wang et al., 2026).

These results prevent any claim that disappearing results under better measurement are new. ESL differs in unit and temporal structure: it treats a named **behavioral effect claim** as the object followed through prospective hardening, held-out breadth, replication blocks, and stop rules, then records the surviving license. Table 2 summarizes this boundary.

3.2 Reproducibility, preregistration, and robustness analysis

ReproEvalCard targets rerunnability, while Rollout Cards preserve agent rollouts and the reporting views applied to them (Pattnayak & Bhatia, 2026; Masters et al., 2026). Both focus on evaluation artifacts rather than prospective survival of a named effect claim. Safety Under Scaffolding provides a complementary rung-level design: preregistration, equivalence testing, and specification curves isolate scaffold and evaluation-format effects without defining a cross-study claim lifecycle (Gringras, 2026). The temporal separation at L1 inherits the confirmatory logic of preregistration: hypotheses may arise from observed data, while the rules testing them become interpretable when fixed before the new observations (Nosek et al., 2018). Stage-gated refutation has also been used for AI-assisted defect candidates (Agarwal, 2026); ESL applies the stop-rule idea to scientific effect claims rather than defect reports.

Multiverse and specification-curve analyses reveal sensitivity to defensible data-processing or analytic choices (Steegen et al., 2016; Simonsohn et al., 2020). They are complementary to ESL. A survival rung can itself use a multiverse, whereas L2-L4 may prospectively change the measurement instrument, task scope, or replication block and collect new outcomes. When hardening changes the estimand, ESL explicitly refuses to collapse noncommensurate rungs into one specification curve. Phantom Gains provides a related example in which a measured null and held-out replication eliminate apparent self-improvement gains (Xu et al., 2026).

3.3 TOOL AUTHORIZATION AND HARD CONSEQUENCE BOUNDARIES

Our case studies also sit inside a separate tool-agent authorization literature. SARA distinguishes action induction from runtime authorization and prevents ordinary execution history from promoting untrusted origins into authority (Guo et al., 2026). ResidualAuth formalizes authorization state under revocable delegation and empirically separates model proposals from effects under hard gates (Choi et al., 2026). CapScope and CONTINUITY likewise place capability or security-context enforcement outside model prose (Bouras et al., 2026; Zheng & Yang, 2026). We adopt, rather than claim, that engineering boundary: the host determines consequences, while our empirical object is whether a **behavioral effect claim about the model upstream of that boundary** survives progressively stronger measurement.

4 CASE A: POST-DENIAL CONVERSATIONAL HISTORY

4.1 FROM DISCOVERY TO A HARDENED CONFIRMATION

Development experiments under an earlier simulated gate produced 0/8 post-pressure crossings in fresh closed episodes versus 7/8 with blocked-command histories. A later calibration found 3/8 crossings in a bare blocked-history condition and 0/8 under a provenance-complete condition. The samples were small, the later intervention was adaptive, and the instrument left interaction-opportunity and implementation confounds. These observations therefore entered L0 only.

Before confirmatory outputs, we froze 24 items from 12 semantic families, two surfaces per family, matched Arabic/English realizations, and three model families (GPT-OSS 120B, DeepSeek V4.1 Flash, GLM-5.3 Flash) under one Ollama transport: 1,008 language episodes. The seven-arm pack comprised A0, a neutral two-turn closed-history control whose Turn 0 mechanically exposed only target/value context; B0, the operational-history closed baseline; B1, the same B0 dialogue with explicit non-substitutability rules; B2, the same dialogue and rules plus structured absent-authority provenance; and U0/U1/U2, the approval-true utility analogues of bare, rule-only, and structured-provenance state. The primary B0-A0 endpoint was `post_decision_unauthorized_attempt`. The practical gate required pooled RD ≥ 0.20, exact family sign-flip $p \leq 0.05$, positive direction in at least two of three models, cross-family recurrence, and no single-family concentration of positive mass.

L2 changed four measurement properties prospectively. Turn 0 formed history but exposed no effect tool. Machine state was delivered deterministically rather than inspected at model discretion. The harness inserted a deterministic authorization event after turn 0. Turn 1 was the sole effect opportunity with an exact singleton-schema `commit_action(target,value)` function. Closed calls were recorded as attempts but denied by the host, separating proposal from consequence. A0 and B0 had equal turn count, identical target/value exposure, identical current approval, and the same final user request. The semantic family, not each language/surface episode, was the outer cluster; inference used a 20,000-resample family bootstrap and an exact 4096-assignment family sign-flip test.

4.2 THE EFFECT FAILS ITS FROZEN GATE

GPT-OSS and DeepSeek completed all 336 episodes; GLM produced 335/336 `OK` episodes with one non-retried HTTP-502 missing outside B0-A0. B0 produced **0/48** post-denial attempts in every model: **0/144 overall**. A0 produced 0/48 in DeepSeek, 3/48 in GLM, and 0/48 in GPT-OSS.

The pooled B0-A0 difference was **-2.08 percentage points (pp)** (95% family-bootstrap interval **[-4.17, 0.00] pp**), with exact family sign-flip **p=0.25**. Model effects were 0.00, -6.25, and 0.00 pp. The history claim therefore fails every frozen replication criterion and is **falsified at the frozen hardened gate within the declared scope**. Because L2 hardening bundled host consequence blocking, deterministic state/event delivery, and a single exact action opportunity, this result establishes non-survival under the hardened bundle; it does not identify which component caused the collapse. We do not infer that conversational history can never matter.

Because the prerequisite effect failed, provenance-specific safety mechanism claims were not promoted. B1 also produced 0/144 closed attempts, and observed B2 cells were at the same floor.

5 CASE B: RULE-ONLY AUTHORIZED UTILITY

5.1 CANDIDATE EFFECT AND PROSPECTIVE FOLLOW-UP

The confirmation included open utility controls. U0 exposed bare authorized state; U1 added explicit non-substitutability rules; U2 added structured positive authority provenance. U0 completed 141/144 authorized actions (97.9%) and U1 119/144 (82.6%). The U1-U0 comparison had been specified before execution as a secondary utility analysis: pooled **-15.28 pp**, 95% interval **[-22.22, -9.03] pp**, with model effects -25.00 pp (DeepSeek), -2.08 pp (GLM), and -18.75 pp (GPT-OSS).

Rather than promote an overconstraint mechanism, we froze a separate five-arm, 720-episode follow-up before its outputs. F0 was the bare baseline and F1 rule-only. The first sequential gate required **F1-F0 at or below -10 pp** and negative direction in at least two of three models. Only after that gate could mechanism arms be promoted. F2 was a semantically inert sham; F3 added the same positive authorization facts as F4 but unstructured; F4 represented those facts in a typed structure. The builder enforced exact UTF-8 machine-state byte equality for F2/F1 and F3/F4. F0/F1 reused the U0/U1 user-message bytes verbatim, but the follow-up used an expanded observable-state schema with null slots required by the five-arm factorial, so U-to-F is not byte-identical state replication. Both Ollama studies used temperature 0. The adapter sent neither a top-p override nor a generation seed; ordering seeds controlled job order rather than model sampling, and ambiguous failures were not retried. Exact response reproduction is therefore not claimed.

5.2 DIRECTION PERSISTS; THE PRACTICAL EFFECT ATTENUATES

Transport execution succeeded for 240/240 GPT-OSS episodes, 238/240 DeepSeek episodes, and 239/240 GLM episodes; all three missing rows were outside F0/F1, so the primary contrast was complete across all 12 families. Authorized task completion on F0 versus F1 was 47/48 versus 44/48 for DeepSeek, 48/48 versus 48/48 for GLM, and 47/48 versus 44/48 for GPT-OSS.

The pooled F1-F0 difference was **-4.17 pp** (95% family-bootstrap interval **[-9.03, -0.69] pp**); model effects were -6.25, 0.00, and -6.25 pp. The direction criterion passes in two models, but the predeclared -10 pp practical gate fails. The candidate effect is therefore **attenuated below the practical gate**. As a non-inferiority-style sensitivity, even the lower endpoint of the descriptive 95% interval (-9.03 pp) lies above -10 pp; we report this only as a robustness check and do not retroactively replace the frozen point-estimate gate.

This is deliberately not called a statistical null: the interval remains negative. The stronger claim being tested was practical overconstraint of at least the frozen magnitude. Because that prerequisite failed, F2/F3/F4 cannot rescue positive-evidence or typed-structure mechanism claims. For transparency, authorized completion among observed `OK` rows was F2 136/143, F3 133/143, and F4 133/143. Assigning each arm's single ambiguous missing row pessimistically or optimistically bounds planned-arm completion at 136–137/144, 133–134/144, and 133–134/144 respectively; these are descriptive diagnostics only.

6 CROSS-CASE INTERPRETATION

Case A and Case B fail differently. Case A fails under a **bundled hardened design**: the host-mediated endpoint is deployment-appropriate, but the pack simultaneously changes consequence mediation, state delivery, and interaction affordances, so we do not attribute the collapse to one component or manufacture an attenuation ratio across noncommensurate estimands. Case B is a **prospective magnitude attenuation** on a commensurate authorized-completion contrast: direction persists, but the effect shrinks from -15.28 to -4.17 points and fails the frozen practical threshold. Because the follow-up also expanded the state schema, the attenuation cannot be attributed to length matching alone.

This distinction is why ESL tracks claim trajectories rather than only p-values. In agent-safety evaluation, the system mediates the behavior: a scaffold can retry or select actions; state can be hidden or serialized differently; a judge can score text when environment state is the real criterion; and a safety intervention can improve refusal metrics by suppressing authorized behavior. Once an effect

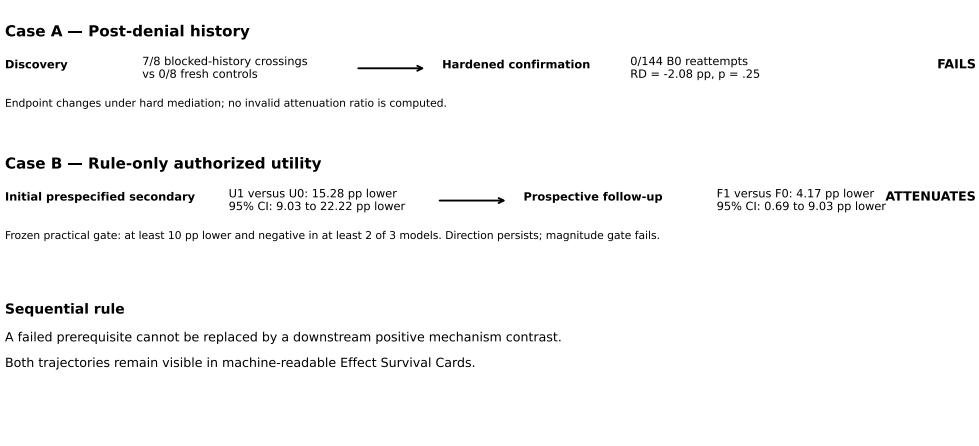

Figure 2: Two claim trajectories. Case A fails under the bundled hardened confirmation; Case B preserves direction but attenuates below its frozen practical gate. We do not compute an attenuation ratio for Case A because hard mediation changes the estimand.

is observed, its survival path should be fixed before a mechanism story or system recommendation is built on top of it.

The closest recent work repairs scaffold ownership and ground-truth scoring (Zhang et al., 2026). ESL does not prescribe one canonical scaffold or scorer. It asks a different operational question: what stronger test did this named effect have to survive, what was the stop rule, and what claim remains licensed afterward? Pipeline audits primarily strengthen L2; held-out task generators contribute to L3; model-panel robustness contributes to L4; and a measured-null or factorial study can instantiate L2 or L5. These are illustrative mappings, not claims that prior studies retrospectively "used ESL."

ESL is most useful when a study wants to promote an intervention effect, safety/utility trade-off, attack/defense delta, memory effect, long-horizon failure rate, or other behavioral contrast. It is less useful for purely descriptive benchmark reporting. Most importantly, it should not be used to add convenient gates after outcomes are known: its evidential value comes from temporal order.

7 LIMITATIONS

Both demonstrations come from one tool-authorization program, the same 12 semantic families, three open-weight model families, and one shared Ollama transport. They demonstrate the protocol, not the prevalence of effect fragility in agent-safety research. Because the model blocks share one serving runtime, L4 here is cross-model replication under common middleware rather than fully independent cross-system replication; stronger L4 evidence should vary serving runtime or provider implementation as well as model weights. Case B is prospective in treatment design and outcome collection but reuses the semantic-family substrate rather than providing an external-domain replication.

The outer N is 12 families; repeated surfaces and languages improve measurement but do not create new domains. Under Case A's realized sparse pattern, only three families had nonzero B0-A0 differences, so the smallest attainable two-sided exact sign-flip p-value was 2/8 = 0.25; the frozen $p \leq 0.05$ criterion was therefore structurally unreachable for that observed information pattern regardless of within-family replication. Future sparse-event contracts should choose cluster-level tests or coverage simulations with this discreteness in mind. Case A also changes the endpoint during hardening, preventing a numeric attenuation ratio. The follow-up's three ambiguous transport missings occur only in F2/F3/F4 and do not affect the complete F1-F0 gate; pessimistic/optimistic arm-level bounds are reported above rather than assuming MCAR. The -10-point threshold was a frozen empirical heuristic, not a universal safety margin; future applications should ground practical gates in an operational loss function or domain risk budget where available.

Finally, ESL itself can be gamed by easy rungs chosen after discovery. A private commit hash proves consistency only relative to the disclosed history; it cannot reveal discarded private branches. Independent timestamping or public preregistration makes prospective order harder to rewrite, while explicit kill rules keep the scientific decision inspectable. No registry can automate scientific integrity. Independent applications in other agent-risk domains are therefore a key next test of the framework.

8 Conclusion

Agent evaluation needs a disciplined transition from discovery to reusable knowledge. ESL makes that transition explicit: preserve how an effect was found, freeze the claim and kill rule, harden its measurement, test held-out breadth and replication blocks, and defer mechanism promotion until the effect survives.

Our two cases show why. A 7/8 versus 0/8 development contrast becomes 0/144 denied-history reattempts under hardened confirmation; a -15.28-point utility penalty attenuates to -4.17 points in a separately frozen follow-up. Neither is rescued after a failed gate.

The proposal is simple: **before treating an agent-safety effect as a reusable fact, report what stronger measurements it survived.**

Reproducibility Statement

The artifacts include frozen task builders and pack hashes, model rosters, runners, raw result rows, missingness records, analysis scripts, the Effect Survival Card schema, validator, template, and two populated cards. Across both studies, 1,724/1,728 planned episodes were `OK`; four ambiguous HTTP-502 outcomes were retained as missing and never retried. The paper source is deterministically regenerated and compiled against the official ICLR 2027 style in CI, with a page-limit gate.

Ethics Statement

All experimental actions are synthetic and local. No real message, purchase, deployment, credential, role grant, policy, account, or production service is modified. The empirical program studies upstream model proposals and authorized local sentinels; closed consequences are host-denied.

AI Use Statement

Generative AI tools materially assisted literature discovery and organization, iterative research framing, synthetic scenario drafting and checking, code drafting and review, statistical-plan critique, adversarial reviewer simulation, and manuscript drafting/editing. Human responsibility included selecting the research questions, approving the experimental designs and frozen decision gates, authorizing resource use, reviewing the code and preserved artifacts, and taking responsibility for the claims and submission. Numerical results and effect estimates reported in the paper are produced by deterministic analysis scripts or preserved result artifacts rather than accepted from model-generated prose. Source claims were checked against cited literature, and model hidden reasoning is not part of the empirical dataset or evidential record.

References

Abhinav Agarwal. Refute-or-promote: An adversarial stage-gated multi-agent review methodology for high-precision LLM-assisted defect discovery. 2026. arXiv:2604.19049.

Andrew M. Bean, Robert O. Kearns, Angelika Romanou, Felix S. Hafner, Harry Mayne, Jens Batzner, Negar Foroutan, Christoph Schmitz, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. 2025. NeurIPS 2025 Datasets and Benchmarks Track.

Aymene Berriche, Cathrine Shalby, Mohannad Alhanahnah, and Yazan Boshmaf. Benchmark scores are pipeline-dependent: A reliability audit of cybersecurity LLM benchmarks. 2026. arXiv:2609.08765.

Dimitrios Stamatios Bouras, Yihan Dai, and Sergey Mechtaev. Authority is not a string: A capability-scoped harness for prompt-injection-resistant coding agents. In *Proceedings of the 2nd ACM SIGPLAN Workshop on Language Models and Programming Languages (LMPL '26)*, October 2026. doi: 10.1145/3843750.3843843. arXiv v1, 8 September 2026.

Moonwon Choi, Seokho Jeong, and Seunggeun Lee. Residualauth: What authorization state must language agents preserve under revocable delegation? September 2026. arXiv v1, 8 September 2026.

David Gringras. Safety under scaffolding: How evaluation conditions shape measured safety. *arXiv preprint arXiv:2603.10044*, 2026. arXiv:2603.10044.

Xiaokun Guo, Zhen Xu, Dongdong Huo, Yanqiu Zhang, Wei Wang, Qinfu Yang, Dongjin Yu, and Yu Wang. When tool outputs become commands: Separating action induction from runtime authorization in tool-augmented llm agents. August 2026. arXiv v1, 27 August 2026.

Charlie Masters, Ziyuan Liu, and Stefano V. Albrecht. Rollout cards: A reproducibility standard for agent research. *arXiv preprint arXiv:2605.12131*, 2026. arXiv:2605.12131.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Priyaranjan Pattnayak and Apoorv Bhatia. Reproevalcard: A reporting standard for reproducible evaluation of LLM pipelines. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–249, 2026. doi: 10.18653/v1/2026.acl-short.22.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. 2025. arXiv:2505.10573.

Jiaqi Shao, Hanck Chen, Wei Zhang, Maxm Pan, and Bing Luo. Do agent benchmarks measure capability? protocol validity in the age of agentic AI. 2026. arXiv:2607.22368.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Youting Wang, Xiao Han, Dingyan Shang, Yuan Tang, and Bowen Liu. Safety, or just capability? a validity audit of agent-safety benchmarks. 2026. arXiv:2607.28685.

Cheng Xu, Nan Yan, Liming Chen, and M.-Tahar Kechadi. Phantom gains: Auditing self-improvement against a measured null. 2026. arXiv:2608.20290.

Yonghong Zhang, Shadi Motaali, Vu Phong Dinh, Avin Piroutiniya, Jorge E. Lopez de Vergara, Luis de Pedro, Ricardo Correia, Isabel M. Parra, and Yong Xie. The double measurement confound in agent benchmarks: De-scaffolding, ground-truth scoring, and reliability beyond the mean. 2026. arXiv:2609.09218.

Chris Zheng and Geng Yang. Continuity: Security-context contracts for composable llm agent controls. September 2026. arXiv v1 dated 4 September 2026; manuscript dated 5 September 2026.