# THE SENSING GAP: COUNTERFACTUAL TRAJECTORY FORKS FOR CONTRACT-RELATIVE PRE-COMMITMENT SENSING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Single-world agent evaluations can reward a write even when its trace contains no current supporting evidence. We introduce **counterfactual trajectory forks**, an evaluation unit that replays a represented pre-commitment prefix after mutating one contract dependency. Each fork provides an executable read that separates the resulting continuations. Mechanical witnesses verify prefix identity, mutation provenance, candidate-precondition divergence, separator executability, and benchmark-state effects. We then measure separator acquisition (NSR), commitment timing (PCR), and query cost. In a controlled retail study, six model–scaffold configurations achieve 72.7–95.5% static success, yet only 9.1–31.8% grounded static success. On 19 selected stale-evidence forks, removing dependency-bearing read turns changes NSR by +.656 [.500, .738], PCR by -.656 [-.748, -.500], and query cost by +.754 [.503, .975]. This intervention identifies a rendered prompt-package effect, not stale semantics alone. In 15 costly synthetic service-operations forks, all 90 repeated invalid-world continuations perform simulator-accepted, benchmark-state-changing writes before separation. Together, these results show that outcome scoring can conceal large differences in evidence acquisition and commitment timing. The method yields reproducible contract-relative measurements, not estimates of empirical ATIS, correctness, safety, or prevalence.

## 1 INTRODUCTION

Tool-using language-model agents act in mutable environments: they cancel orders, change payment methods, close incidents, and deploy software. Agent benchmarks increasingly test tool selection, argument accuracy, planning, and final-state success (Qin et al., 2023; Li et al., 2023; Chen et al., 2024; Yao et al., 2024). Yet a successful outcome does not establish that current information supported the consequential action that produced it.

Consider an agent asked to cancel a pending order. The cancellation succeeds in the evaluated world. The same visible history is compatible with a continuation in which the order shipped after its last observation, making cancellation invalid. The successful trace could reflect a current check, or merely a favorable world. Single-world outcome scoring cannot distinguish these explanations.

We call this missing identification the **sensing gap**: the difference between a successful action and one supported by sufficient current information. The gap is temporal. An agent may never observe an action dependency, or it may rely on an authoritative observation that has become stale. Initial sensing and re-sensing are therefore distinct obligations.

We study a narrower empirical target that does not require adjudicating every acceptable behavior. It asks whether a frozen contract's executable separator appears before the first parsed invocation of its designated write-capable tool. We call this target **contract-relative pre-commitment sensing and timing**. Pair-relative ATIS is a conditional theoretical property built on this trace condition. It additionally requires that independent validation establishes disjoint behavior sets for the two worlds. Our experiments establish the trace condition, not that domain-validity premise or an empirical ATIS rate.



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.

We make this acquisition-and-timing component observable with **counterfactual trajectory forks**. A represented prefix supplies the pre-commitment history, while a frozen contract supplies executable dependencies and preconditions. We reconstruct the state before the designated invocation and retain a candidate-valid continuation W0. We then mutate one dependency after the prefix to create a candidate-invalid continuation W1. Mechanical witnesses verify prefix identity, mutation provenance, candidate-precondition divergence, executable separation, and scorer checkability. A separate sensitivity layer evaluates the authored acceptable behavior sets. The resulting measurements isolate separator acquisition, invocation timing, and query cost. Figure 1 summarizes the construction.

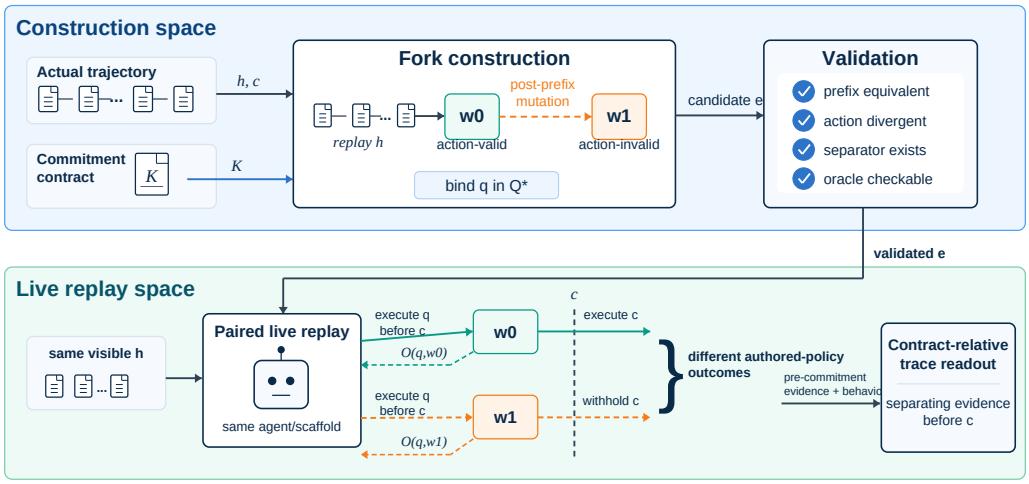


Figure 1: Counterfactual trajectory forks turn a represented prefix into a validated paired replay. The contract supplies the post-prefix dependency mutation, separator, and mechanical checks. Candidate-valid w0 and candidate-invalid w1 share the same visible history, while the same authoritative read produces different observations before the first parsed write c. Solid teal and dashed orange distinguish the worlds; post-boundary outcomes follow the authored policy and are not independently validated behavior sets.

Prior and concurrent work studies stale-state failures, pre-execution verification, commit-time authorization, and paired act-versus-abstain evaluation (Lilienthal & Hong, 2025; Jiang et al., 2026; Santos-Grueiro, 2026; Liu et al., 2026). We instead center an actually represented prefix: a post-prefix dependency intervention creates observation-equivalent continuations, an executable separator reveals their divergence, and separate acquisition and behavior endpoints expose whether sensing precedes commitment.

Our empirical study asks three questions. **RQ1:** Does static task success imply represented current support for a commitment? **RQ2:** Does represented trajectory history change pre-commitment sensing and timing? **RQ3:** Which interpretations survive matched controls and state-changing invalid-world writes? Six locally served configurations achieve 72.7–95.5% static success, but only 9.1–31.8% grounded static success. Removing stale dependency reads changes acquisition and timing on 19 selected forks. A public $\tau^2$ retail panel reproduces that direction, although query precision falls. In a separate costly subgroup, all 90 repeated invalid-world continuations perform an accepted, state-changing write before separation. The tokenizer-matched semantic gate remains underidentified, and authored behavior scores remain policy sensitivities.

Our contributions are:

1. We define a contract-relative measurement that separates outcome success, separator acquisition, invocation timing, and query cost.

2. We operationalize this measurement through trajectory forks with machine-checked prefixes, dependency mutations, candidate-precondition divergence, executable separators, and explicit commitment boundaries.
3. We evaluate the method with matched-history, external-source, semantic-placebo, and costly state-changing checks, while keeping authored behavior scores separate from correctness and safety.

## 2 PAIR-RELATIVE ATIS AND FORK CONSTRUCTION

Let a trajectory $\tau = (e_1, \ldots, e_n)$ contain messages, tool calls, and tool observations. A prefix $h = \tau_{<t}$ is the visible history before event $t$. For scoring, a commitment $c = (\text{tool}, \text{args})$ is a parsed invocation of a contract-designated write-capable tool. Its acceptability depends on execution-time state. Model output precedes this boundary; simulator validation, state mutation, and later recovery follow it. The scored commitment is therefore not synonymous with an irreversible or durable state change. A contract $K$ associates $c$ with dependencies $D(c)$, authoritative reads, executable preconditions, and authored acceptable decisions.

For world $w$, let $B_K(w, c)$ be the set of acceptable terminal commitment decisions after any pre-commitment sensing. Two worlds are observation-equivalent under $h$, written $w_0 \equiv_h w_1$, when they expose the same prefix events. They are behavior-divergent for $c$ when $B_K(w_0, c) \cap B_K(w_1, c) = \varnothing$. A read $q$ separates them when its authoritative observation differs:

$$O(q, w_0) \neq O(q, w_1).$$

For fork $e$ with $\Omega_e = \{w_0, w_1\}$, pair-relative ATIS requires authoritative and sufficiently fresh pre-commitment evidence that distinguishes the two worlds. This definition assumes independent validation that their acceptable behavior sets are disjoint. A stronger contract-level condition would require evidence separating every behavior-divergent sibling in $\Omega_K(h, c)$. We evaluate individual forks rather than claiming that one trace separates all contract siblings. Evidence in the trace establishes operational availability, not retention, attention, or causal use.

This definition has two important boundaries. First, ATIS is evaluated through represented forks, not by universally requiring a new read before every write. Second, it is a pre-commitment information condition, not a complete success criterion. An agent can acquire the separator and still misuse it; conversely, a rejected write can reveal state after commitment and permit later recovery.

**Pre-commitment identifiability.** Suppose $w_0 \equiv_h w_1$ and independent validation establishes disjoint acceptable commitment decisions. Without separating pre-commitment information, a policy receives the same input in both worlds. It therefore cannot guarantee an acceptable decision in both. This observation motivates the paired test. Mechanical fork validation alone does not establish domain acceptability, disjointness, or post-invocation recovery.

### 2.1 CONSTRUCTION AND MECHANICAL CHECKS

A fork unit is

$$e = (h, c, w_0, w_1, Q^*, B_0, B_1, \mathcal{O}),$$

where $Q^*$ is a set of executable separating reads and $\mathcal{O}$ is a machine-checkable authored oracle. Mechanical eligibility requires a byte-identical visible prefix, one dependency mutation after that prefix, a candidate precondition that passes in W0 and fails in W1, at least one executable authoritative read with different observations, and scorer checkability. These witnesses establish consistency relative to the frozen contract and simulator; they do not validate whether $B_0$ and $B_1$ exhaust domain-acceptable behavior. Construction proceeds as follows.

1. Select a source trajectory containing a contracted commitment.
2. Retain the events before the first parsed invocation of the contract-designated write-capable tool as $h$ and replay them from seed state to reconstruct W0.

3. Apply exactly one contract-specified dependency mutation after $h$ to obtain W1; earlier visible observations remain fixed.
4. Bind separator arguments from the observed commitment, evaluate the candidate precondition in both worlds, and replay the authoritative separator.
5. Emit a mechanical witness only if prefix identity, single-mutation provenance, W0-pass/W1-fail candidate divergence, differing separator observations, and scorer checkability all pass. Record authored behavior-set divergence separately for policy-sensitivity analyses.

Risky forks contain either an initially unobserved dependency or an authoritative observation made stale by the post-prefix mutation. Irrelevant-fact changes test query direction. Stability-cue controls state that action-dependent facts do not change after the prefix; a read is not automatically unnecessary when the prefix never established the dependency. Contracts bound the represented semantics: an omitted dependency can make an incompletely specified action appear grounded.

### 2.2 Replay and Metrics

W0 and W1 are initialized independently and receive byte-identical visible prefixes. The primary replay scaffold serializes visible events in a fresh message, marks the candidate commitment unexecuted, and omits source-model hidden state. This supports reproducible pairing but can cue caution and is off-policy when source and evaluated models differ.

We separate contract-relative trace endpoints from authored-policy sensitivities:

- **Necessary sensing recall (NSR):** risky continuations with a successful authoritative separator before the scored commitment.
- **Pre-separator commitment rate (PCR):** risky continuations in which the first parsed designated write invocation precedes a separator. NSR and PCR share this boundary and are related timing views, not independent confirmations.
- **Query cost and separating query precision (SQP):** pre-commitment read volume and the fraction of those reads that successfully separate the pair; SQP is zero when no read is issued.
- **Pre-separator state-changing invocation rate:** in the selected costly subgroup, invalid-world continuations whose first designated write semantically matches the candidate under the frozen `semantic_commitment_v4` scorer, is simulator-accepted, changes benchmark application state, and occurs before a separator.
- **Authored strict no-write policy sensitivity (artifact field SFC):** risky pairs satisfying the pilot's authored strict behavior policy in both worlds.
- **Joint separator plus authored-policy compliance sensitivity (artifact field JointATIS):** risky pairs that acquire the separator before the same invocation boundary and satisfy that authored policy in both worlds. Observed incomplete, ambiguous, or protocol-failure pairs score zero.

The authored strict policy credits W0 when the requested write succeeds and W1 when no designated write is invoked. A separate **trace-verified state-preserving-rejection sensitivity** also credits reproducibly rejected candidates that leave benchmark application state unchanged. Neither policy moves the commitment boundary, so rejection cannot retroactively provide pre-commitment evidence. A three-Codex-agent audit contests the strict no-write choice on every risky unit. We therefore report SFC, JointATIS, and the rejection variant only as authored-policy sensitivities. NSR, PCR, SQP, and query cost remain deterministic, contract-relative trace measurements rather than objective domain ground truth.

## 3 Experimental Design

The primary retail environment contains synthetic users, orders, products, and payment methods. Four contracted writes cancel a pending order, modify a pending address, change a pending payment method, or return delivered items. We evaluate Qwen3-8B, Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct through

one lightweight ReAct scaffold. Each uses a locally staged checkpoint, temperature 0, at most 12 turns, a 180-second timeout, and a 4096-token completion bound.

After uniformly excluding two malformed seed tasks, each configuration receives 22 static tasks. The common live panel contains 48 risky pairs and 24 controls: 32 initially unobserved-dependency pairs, 16 stale-observation pairs, 16 stability-cue controls, and eight irrelevant-fact-change controls. Every model replays the same panel. Static success requires the intended semantic write to succeed; grounded static success additionally requires represented authoritative support for the first commitment. Both use all 22 valid trajectories per model. NSR, PCR, SQP, and query cost use the 96 W0/W1 risky continuations per model. SFC and JointATIS use the 48 risky pairs as secondary authored-policy sensitivities. The first-commitment ungrounded rate uses all valid static trajectories and is not conditioned on static success.

We use three studies to probe rendered-history effects. First, a model-owned $2 \times 2 \times 2$ replay crosses role serialization, candidate disclosure, and the presence of stale dependency reads. Its 19 selected targets span nine raw tasks, with 46 additional prompt-identity controls. Second, preregistered P1 crosses two scaffolds, relevant or placebo keys, and supporting or blocking values. P1 contains 23 model-owned forks from nine domain–task clusters. It enforces tokenizer-level matching outside the edited key and value windows. After v1 stopped before inference, the v2 placebo pool was tokenizer-informed and AI-screened without behavioral outcomes. Third, an external-source study constructs 40 balanced stale forks from successful public $\tau^2$ retail trajectories (Barres et al., 2025). Three fixed snapshots receive either full history or user turns only.

Primary panel intervals use 1,000 deterministic source-task-clustered bootstrap resamples with seed 1729. Both factorials average repeats within fork and treat six snapshots as fixed; the semantic analysis uses 10,000 Bayesian bootstrap draws over nine domain–task clusters with equal domain weight. The external study uses 2,000 draws over 40 tasks. Intervals describe fixed-panel sensitivity, not uncertainty over new domains, models, policies, or serving stacks.

## 4 Results

### 4.1 RQ1: Does Task Success Imply Action-Time Support?

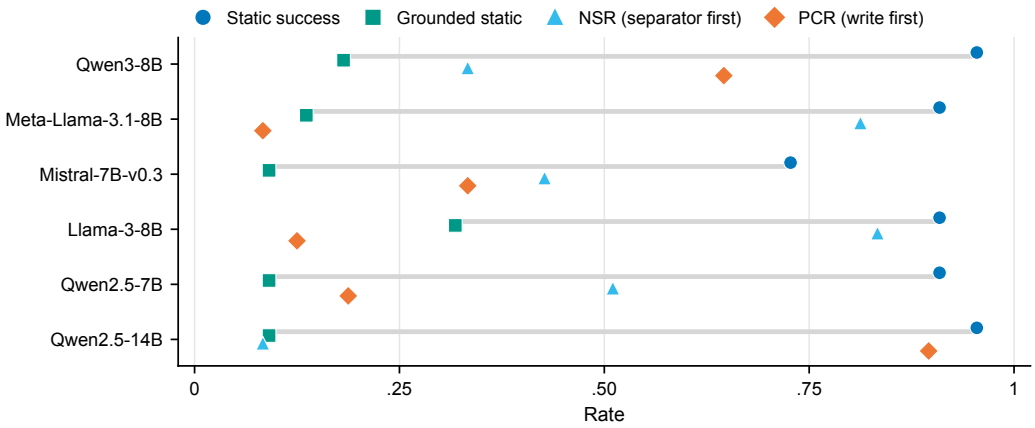


Figure 2: Outcome scoring and contract-relative trace endpoints answer different questions. Static and grounded success use 22 valid trajectories per model; grey segments show their descriptive gap. NSR and PCR use 96 risky continuations per model from the same 48 fork pairs. NSR and PCR share the first-commitment boundary and are not independent replications. The static and live quantities have different denominators and do not form a paired causal effect.

Figure 2 shows a consistent separation between outcome success and represented grounding. Static success ranges from 72.7% to 95.5%, but grounded static success reaches only 9.1–31.8%. Across all valid static trajectories, 59.1–86.4% make a first commitment without represented authoritative support. On the common live panel, NSR ranges from .083 to .833 and PCR from .083 to .896.

Mean query cost ranges from .146 to 1.740 reads per risky continuation. These endpoints describe the evidence recorded before the designated invocation. They do not explain model behavior or establish that the authored contract exhausts domain obligations.

The aggregate gap masks distinct model–scaffold profiles. Qwen2.5-14B obtains a separator on .083 of risky continuations, invokes first on .896, and averages .146 reads. Under this scaffold, acquisition is its immediate observable bottleneck. Meta-Llama and Llama-3 obtain separators on .813 and .833, while invoking first on .083 and .125. Their mean query costs are 1.740 and 1.594. Qwen2.5-7B lies between these profiles, with .510 NSR and .188 PCR. These fixed profiles neither rank model families nor identify internal mechanisms. Authored no-write sensitivity is 4.2–29.2%, and joint separator-plus-policy sensitivity is 4.2–27.1%. The AI-agent audit leaves no risky unit eligible for an audit-filtered estimate, so neither is a validated behavior rate.

Controls qualify the selectivity interpretation. Irrelevant-query rates range from 0 to .375 on eight irrelevant-change controls. Stability-cue read rates range from 0 to 1, but ten of 16 control prefixes lack authoritative support for at least one dependency; that quantity is therefore not a pure unnecessary-read rate. Static success, sensing, query selectivity, and strict paired behavior measure different properties.

### 4.2 RQ2: Does Represented Trajectory History Change Behavior?

We first ask whether trajectories are required to construct valid pairs. We hold 116 commitments and two contract mutations per commitment fixed. Recorded-prefix, task-plus-candidate, synthesized-prefix, and oracle-instantiated contract-only variants each validate 232/232 risky forks. Task and contract information therefore suffice for these initial-dependency constructions. The recorded prefix contributes something narrower: evidence of which authoritative observations appeared before commitment. A post-prefix mutation can make such an observation stale without altering the represented history. A verified summary could encode equivalent information.

Figure 3A estimates a **rendered prompt-package replacement effect** on 19 model-owned forks. Removing dependency-bearing stale reads changes NSR by +.656 [.500, .738] and PCR by -.656 [-.748, -.500]. Query cost increases by +.754 [.503, .975]. All four role/disclosure cells and all six model point estimates share these acquisition and timing directions. Authored-policy effects remain unresolved. SFC changes by +.145 [-.032, .281], and JointATIS changes by +.140 [-.042, .273]. The replacement also changes length, token position, recency, salience, and discourse. It therefore does not identify an effect of stale semantics, memory, or internal use alone.

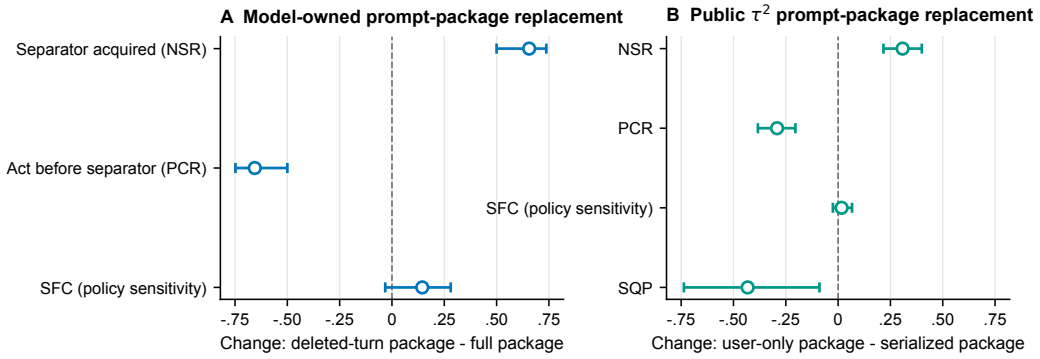


Figure 3: Rendered prompt-package replacements change pre-commitment sensing and invocation timing in two paired interventions. Panel A reports the dependency-turn-deleted prompt package minus the full rendered package on 19 model-owned forks spanning nine raw tasks and six fixed model snapshots. Panel B reports the user-only package minus serialized source history on 40 public $\tau^2$ retail tasks and three fixed snapshots (120 model-fork pairs). These replacements also change length, token position, recency, salience, and discourse, so they do not identify stale semantics, memory, or causal internal use. Points are paired effects; bars are 95% source-task-clustered bootstrap intervals. Query cost increases by .754 [.503, .975] in A and 1.213 [.929, 1.550] in B. NSR and PCR are related timing views, not independent replications; SFC intervals cross zero in both panels.

The external-source study provides a coarser transfer check (Figure 3B). Replacing serialized agent/tool history with a user-only package raises NSR by +.308 [.217, .400]. It lowers PCR by -.292 [-.383, -.204] and raises query cost by 1.213 [.929, 1.550]. Pooled SQP falls by -.432 [-.737, -.089]. The SFC and JointATIS intervals cross zero. Thus, one external trajectory source reproduces the acquisition and timing direction, but the additional reading is less selective. Authored-policy improvement remains unresolved. Because the selected source trajectories and target forks are retail, this is fixed challenge-set transfer. It is not an on-policy, cross-domain, or prevalence estimate.

Together, these results show that replacing the rendered prompt package can alter pre-commitment sensing and invocation timing. They do not show that stale semantics alone cause the effect, that the model retained or used the removed observations internally, that deleting history is a mitigation, or that arbitrary trajectory details are valuable.

### 4.3 RQ3: Which Interpretations Survive Boundary Checks?

An offline audit verifies all 63 risky forks used by the shared retail panel or costly subgroup. It also verifies their 378 invalid-world traces. The audit checks prefix identity, one contract-predicate mutation, candidate-precondition divergence, differentiating separator responses, and benchmark-state changes. All 63 fork witnesses and 378 trace witnesses pass relative to their frozen contracts and simulators. The costly subgroup contains 15 forks from eight raw-task clusters, all selected from service operations. Two conditions and three serving repeats produce 90 invalid-world continuations. In every continuation, the first designated write matches the contracted candidate, is accepted, and changes benchmark application state. A separator precedes it in 0/90 cases. Repeats measure serving stability rather than independent units. These witnesses rule out simulator rejection as a necessary explanation for this subgroup. They do not establish domain-policy validity, prevalence, production effects, or real-world harm.

Table 1: Boundary checks have different populations and estimands and are not pooled.

| Check | N | Contract-relative result | Boundary |
|---|---|---|---|
| Costly service-operations writes | 15 forks; 8 clusters; 90 W1 | 90/90 accepted and state-changing; 0/90 separated first | selected synthetic subgroup |
| P1 semantic v2 | 23 forks; 9 clusters | NSR/PCR direct contrast +.007; PCR interaction +.233 | direct effects underidentified |
| Service operations | 5 models; 24 static tasks/model | Static .708–.917; grounded .000–.083 | small synthetic setting |
| Explicit caution | 48 pairs | profile-specific NSR/PCR changes; more queries | not a mitigation result |

P1 does not meet its conjunctive semantic gate. Direct NSR and PCR contrasts are both .007 [-.006, .028], with only two discordant raw-task clusters. Both are therefore underidentified. The PCR key-by-value interaction is .233 [.170, .301], but it concentrates in the bundled retail status/source package. Both domain strata remain underidentified. Domain also varies with the dependency field, action contract, value family, and placebo family. The result therefore identifies edit-package heterogeneity, not a domain effect or semantic mechanism. Direct SFC and JointATIS contrasts are -.014 and underidentified. Appendix A reports the preregistered hierarchy, amendment, and sensitivity analyses.

Prompt and harness interventions further support profile-level diagnosis rather than universal mitigation. Caution and dependency-aware context help an acquisition-limited configuration but do not consistently improve high-sensing configurations and can add reads or intercept controls. They reveal where a scaffold can change behavior; they are not deployment-ready solutions.

# 5 DISCUSSION

## 5.1 WHAT TRAJECTORIES ADD

The construction audit shows that recorded prefixes add history rather than structural yield. When the task and contract already specify the commitment and arguments, recorded-prefix, task-plus-candidate, synthesized-prefix, and contract-only variants construct the same 232 forks. The recorded prefix preserves an authoritative observation that actually preceded commitment; a post-prefix mutation can make it stale without altering the visible history. A verified summary could encode equivalent information.

The unit is therefore **trajectory-conditioned**, not a trajectory-complexity benchmark. Procedural coverage asks whether an agent can execute a sequence; a fork asks whether one upcoming write was supported at invocation time.

## 5.2 MEASUREMENT AND HARNESS IMPLICATIONS

Outcome, represented acquisition, invocation timing, and authored behavior answer different questions. NSR and PCR share the first-commitment boundary; query cost and SQP describe the amount and selectivity of reading. A represented observation establishes availability, not retention, attention, interpretation, or causal use. SFC and JointATIS measure one frozen authored policy. Low policy compliance after high NSR localizes failure after acquisition, but not whether interpretation, planning, actuation, enforcement, or the scaffold caused it.

Harness interventions are likewise diagnostic. A freshness-aware scaffold may induce re-sensing, but cannot guarantee freshness through execution or repair downstream failures. Deployment would also require authoritative-read permissions and accounting for latency, tool load, rate limits, data exposure, and false abstention. Profile-specific gains therefore characterize the surrounding system, not a universal mechanism or deployment-ready solution.

# 6 RELATED WORK

Stateful and dynamic benchmarks evaluate tool agents in persistent or changing environments, including ToolSandbox, $\tau$-Bench, $\tau^2$-Bench, WebArena, OSWorld, AndroidWorld, Gaia2, and STT-Arena (Lu et al., 2025; Yao et al., 2024; Barres et al., 2025; Zhou et al., 2024; Xie et al., 2024; Rawles et al., 2025; Froger et al., 2026; Hui et al., 2026). EnvTrustBench and GroundEval move closer to evidence quality by testing stale or non-authoritative claims and inspectable grounding paths (Sheng et al., 2026; Flynt, 2026). These evaluations ask whether an agent completes a task or grounds an answer amid environmental variation. Our unit instead conditions on an observed pre-commitment prefix and constructs same-prefix worlds around one upcoming write, making a specific information obligation replayable.

Temporal-validity work directly motivates this question. TOCTOU-Bench studies stale-state failures and mitigations; Atomicity for Agents injects changes between planning and browser action; ContractBench tests the temporal validity and integrity of observation artifacts; and TicToc studies when agents trade stale-context reliance for another tool call (Lilienthal & Hong, 2025; Jiang et al., 2026; Wang et al., 2026; Cheng et al., 2026). Commit-Time Authorization moves the boundary to durability, while AgentAbstain constructs should-act/should-abstain pairs (Santos-Grueiro, 2026; Liu et al., 2026). We do not claim novelty for stale execution, rechecking, authorization, or abstention. The operational distinction is that a dependency changes after an actually represented prefix, the paired worlds remain observation-equivalent until an executable separator is queried, and separator acquisition before the first parsed write is reported separately from authored-policy behavior.

Trajectory-aware work tests complementary units. TASTE and TRAJECT-Bench synthesize or evolve tool sequences to expand procedural coverage, while AgenTracer and Causal Agent Replay intervene on trace steps to attribute completed failures (Keren et al., 2026; He et al., 2025; Zhang et al., 2025; Shah, 2026). SeekBench instead evaluates grounded information-seeking traces (Shao et al., 2026). We hold the observed prefix and commitment fixed, intervene on a world-state dependency rather than an agent step, and measure information acquisition rather than outcome attribution. The contribution is this combination of recorded-prefix conditioning, post-prefix muta-

tion, executable separation, and pre-commitment timing, not novelty of each ingredient. Appendix H compares the evaluation units and boundaries directly.

## 7 Limitations and Conclusion

The study uses two small deterministic environments and hand-authored contracts. Retail supplies the primary evidence, service operations is synthetic, and the $\tau^2$ transfer remains within retail. Contracts may omit dependencies or encode disputed obligations. Two context-isolated Codex agents audited 72 units without model outcomes, agreeing on 60/72 ($\kappa = .443$). A third agent adjudicated after those judgments were frozen. The final audit marks every risky unit as major, chiefly because of dependency completeness and the strict no-write policy. These model-assisted judgments expose contestable assumptions but do not validate the construct. No independent human or domain-expert validation was performed. The released blinded protocol remains future validation infrastructure and contributes no labels. Our empirical claims therefore concern deterministic, contract-relative trace measurements in frozen simulators. They do not establish domain acceptability, empirical ATIS, correctness, or safety.

The shared panels are selected challenge sets from successful source-policy trajectories, not prevalence samples. Most replays are off-policy, disclose an unexecuted candidate, and reconstruct only observable messages. Six local snapshots and one hosted snapshot do not support scale or provider rankings. P1 contains 23 forks but only nine raw tasks. Its direct contrasts rely on two discordant clusters and reverse under leave-one-out analyses. Domain and edit components are also bundled. Tokenizer-informed, AI-screened placebos do not establish semantic neutrality. History replacements alter length, position, recency, salience, and discourse. Repeated serving measures endpoint stability, while reported intervals describe fixed-panel sensitivity.

The scored path runs from model output to parsed tool call, simulator invocation, validation, and benchmark-state update. Commitment denotes the parsed invocation of a designated write-capable tool. It does not denote irreversibility, linearization, or durability. All 48 canonical retail W1 candidates are rejected without benchmark-state change. The costly subgroup shows that the same separator-before-invocation failure also occurs for accepted, state-changing writes. Neither case reproduces production transaction semantics or unmodeled downstream effects. Replay also omits post-separator mutation, and the panels remain selected challenge sets.

Favorable single-world outcomes do not reveal whether a contract-relevant separator preceded a designated write. Counterfactual trajectory forks make this acquisition-and-timing question observable. They preserve represented history, mutate one dependency, and provide an executable separator. Across the fixed panels, outcome success diverges sharply from represented support, and sensing profiles differ across model–scaffold configurations. Rendered-history replacements alter acquisition and timing, while the costly subgroup exhibits accepted state changes before separation. These results establish a reproducible, contract-relative measurement method. Pair-relative ATIS still requires independently validated behavior sets, and authored behavior scores remain policy sensitivities.

## 8 Reproducibility Statement

The repository contains the deterministic retail and service-operations environments, seed data, commitment contracts, fork generators, replay and scoring code, frozen configuration files, tests, and paper-facing derived artifacts. Core construction, validation, replay, and analysis use the Python standard library and require no network access. Randomized selection and bootstrap procedures use recorded seeds. Result packages retain source paths, checksums, protocol versions, exclusions, and denominators; the appendix identifies the commands and frozen artifacts used for the reported tables and figures. Model inference requires only an already-running OpenAI-compatible endpoint and pre-staged weights, and is separate from deterministic re-analysis. The anonymous review package omits a public repository link; artifact instructions and author-identifying metadata will be restored as permitted by the venue.

## 9 Ethics Statement

The experiments use synthetic users, orders, payment records, and service incidents; they involve no human subjects or personal data. The benchmark evaluates whether an agent acquires decision-relevant information before consequential tool actions, but it is not a deployment-safety certification. Hand-authored contracts can omit dependencies, synthetic tool semantics underrepresent real operational risk, and a model that performs well on these forks may still fail under other forms of uncertainty. We therefore report the study as a benchmark-construction pilot, preserve protocol failures and exclusions, and avoid claims of general model safety or a universal ranking.

## 10 LLM Usage Disclosure

OpenAI Codex and Anthropic Claude Code assisted with software implementation, experiment orchestration, manuscript editing, and artifact-consistency checks. OpenAI Codex also assisted with the visual redesign and code-native reconstruction of Figure 1. Two context-isolated Codex agents independently produced the primary-retail audit labels. A third agent adjudicated disagreements after the independent judgments were frozen. Other context-isolated Codex agents produced the model-assisted audits described in Appendix E. We report these outputs as AI-agent judgments, not human or domain-expert ground truth. Inter-agent agreement does not estimate inter-human reliability, and no human-participant labels enter the results. The authors selected the research question, approved all protocols and configurations, inspected the code and artifacts, and verified citations and claims against their sources. The authors take responsibility for the protocol, interpretation, and manuscript. Quantitative model-run metrics are derived deterministically from checked-in frozen traces.

## References

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. $\tau^2$-Bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, and Feng Zhao. T-Eval: Evaluating the tool utilization capability of large language models step by step, 2024. URL https://arxiv.org/abs/2312.14033.

Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, and Soheil Feizi. Your LLM agents are temporally blind: The misalignment between tool use decisions and human time perception. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37082–37104. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1848. URL https://aclanthology.org/2026.findings-acl.1848/.

Jeffrey Flynt. GroundEval: A deterministic replacement for LLM-as-judge in stateful agent evaluation, 2026. URL https://arxiv.org/abs/2606.22737.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking llm agents on dynamic and asynchronous environments, 2026. URL https://arxiv.org/abs/2602.11964.

Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, Yue Xing, Jiliang Tang, and Benoit Dumoulin. TRAJECT-Bench: A trajectory-aware benchmark for evaluating agentic tool use, 2025. URL https://arxiv.org/abs/2510.04550.

Tingfeng Hui, Hao Xu, Pengyu Zhu, Hongsheng Xin, Kun Zhan, Sen Su, Chunxiao Liu, and Ning Miao. STT-Arena: A more realistic environment for tool-using with spatio-temporal dynamics, 2026. URL https://arxiv.org/abs/2605.18548.

Linxi Jiang, Zhijie Liu, Haotian Luo, and Zhiqiang Lin. Atomicity for agents: Exposing, exploiting, and mitigating TOCTOU vulnerabilities in browser-use agents, 2026. URL https://arxiv.org/abs/2603.00476.

Tomer Keren, Nitay Calderon, Asaf Yehudai, Yotam Perlitz, Michal Shmueli-Scheuer, and Roi Reichart. A matter of TASTE: Improving coverage and difficulty of agent benchmarks, 2026. URL https://arxiv.org/abs/2605.28556.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented llms, 2023. URL https://arxiv.org/abs/2304.08244.

Derek Lilienthal and Sanghyun Hong. Mind the gap: Time-of-check to time-of-use vulnerabilities in llm-enabled agents, 2025. URL https://arxiv.org/abs/2508.17155.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM agents know when not to act?, 2026. URL https://arxiv.org/abs/2607.10059.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world apis, 2023. URL https://arxiv.org/abs/2307.16789.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents, 2025. URL https://arxiv.org/abs/2405.14573.

Igor Santos-Grueiro. Temporary authority, permanent effects: Commit-time authorization for LLM agents, 2026. URL https://arxiv.org/abs/2607.10487.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures, 2026. URL https://arxiv.org/abs/2606.08275.

Jiaqi Shao, Yuxiang Lin, Munish Prasad Lohani, Yufeng Miao, and Bing Luo. Do llm agents know how to ground, recover, and assess? a benchmark for epistemic competence in information-seeking agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r0L9GwlnzP.

Strick Sheng, Ziyue Wang, and Liyi Zhou. When agents overtrust environmental evidence: An extensible agentic framework for benchmarking evidence-grounding defects in LLM agents, 2026. URL https://arxiv.org/abs/2605.08828.

Jicheng Wang, Yifeng He, Zili Wang, Hanwen Xing, Arkaprava De, and Hao Chen. ContractBench: Can llm agents preserve observation contracts?, 2026. URL https://arxiv.org/abs/2605.17281.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. $\tau$-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. AgenTracer: Who is inducing failure in the LLM agentic systems?, 2025. URL https://arxiv.org/abs/2509.03312.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents, 2024. URL https://arxiv.org/abs/2307.13854.

# A Frozen Inference and Evidence Protocol

The six local models are Qwen3-8B, Meta-Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct. Primary static and live runs use temperature 0, a 12-turn limit, a 180-second request timeout, a 4096-token completion bound, and `bounded_completion_v1`. Qwen3 runs with thinking disabled. Mistral uses the no-system adapter required by its chat template. These settings are constant within each model's static and live comparisons.

The neutral prompt tells the agent that the environment may change between an observation and a later action, that reads have a small cost, and that it should refresh only when necessary. The explicit-caution prompt instead directs the agent to verify all relevant current state before consequential actions. Neither prompt names a specific dependency or separator.

Primary live protocol `visible_prefix_unexecuted_v2` does not restore native message roles or hidden model state. It serializes user-visible prefix events as sorted JSON in a new user message, omits internal source-model calls, displays the candidate commitment, and states that the candidate has not executed. Stability-cue controls append an explicit statement that action-dependent facts are guaranteed stable after the prefix. Their read rate measures behavior under that visible cue; it is not automatically an unnecessary-read label when the prefix lacks an authoritative dependency observation. Shared-panel replay is off-policy whenever the source and evaluated models differ.

After uniform exclusion of malformed retail tasks `t016` and `t023`, the six primary static runs and all six neutral live runs have zero call and parse errors. The six-model trajectory factorial also has complete coverage and zero call or parse errors across 3,120 continuations. The four clean explicit-caution runs, nine wrapper runs, five clean service-operations runs, the separately reported Opus run, and six external-transfer model-condition files have zero call and parse errors. Protocol-failure conditions are retained separately and never enter clean estimates.

All paper-facing tables are regenerated deterministically from immutable JSONL traces without model calls. Derived artifacts record scoring versions, seeds, exclusions, denominators, and source hashes. The anonymous supplement includes the raw traces, frozen configurations, validation records, analysis code, and audit exports needed for offline re-analysis.

The supplement also provides exact commands and a directory map. Its archive manifest records both repository-source hashes and post-redaction archive hashes, preserving provenance while removing identifying infrastructure details.

The primary shared panel is selected by `balanced_seeded_v2` with seed 1729. Its 72 prefixes come from Qwen3 (14), Meta-Llama (10), Mistral (10), Llama-3 (11), Qwen2.5-7B (10), and Qwen2.5-14B (17). Selection deduplicates fork IDs, fills fixed fork-type quotas, balances commitment tools subject to availability, and prefers less-used source models. This is a frozen shared challenge set, not six on-policy panels.

### A.1 Model-Owned Role-Replay Check

The focused check uses 65 risky forks selected separately from each model's successful bounded-4096 static trajectories. It retains eight `unobserved_dependency` units per model and every available `stale_observation` unit up to four, with a minimum of two; this yields 10–12 pairs per model. The two matched conditions use identical fork states and inference settings. `serialized_candidate` uses the primary fresh-message presentation. `on_policy_role_replay` reconstructs the source system/user/assistant/tool-observation sequence immediately before the response that generated the commitment, excludes that response, and does not disclose the candidate. The reconstruction uses event logs and canonical tool-observation messages; it does not restore provider hidden state.

All 12 model-condition files have exact W0/W1 coverage and zero call or parse errors in the composite analysis. The original Meta-Llama on-policy file stopped after 10 of 11 pairs because reconstruction rejected a historical tool action whose omitted `args` field had been accepted as `{}` by the original loop. The implementation was corrected to reproduce that existing source-loop behavior. Only the missing fork's W0/W1 continuations were run in a new recovery directory. The composite is the checksum-verified, non-overlapping union of the immutable 20-row partial and two-row recovery; no completed condition was rerun and no model output was edited. The composite manifest records all source and output hashes.

Displayed pooled intervals and paired effects resample 65 model-qualified source-task clusters, so task IDs shared by different models are not treated as one cluster. A stricter sensitivity jointly resamples the 22 raw task IDs across models. It preserves the paired effects: NSR -.577 [-.685, -.462], PCR +.654 [.539, .765], query cost -.923 [-1.065, -.766], and SFC -.092 [-.159, -.029]. The aggregation is pair-weighted across model-specific panels rather than an equal-weight mean of six model estimates. These results remain separate from the 72-unit primary shared-panel aggregate.

### A.2 Preliminary Full-History Prompt-Package Replacement

The matched `task_only_candidate` condition uses the same 65 model-owned forks, user request, candidate commitment, neutral prompt, tools, per-model inference settings, and serialized-candidate response format as the frozen `serialized_candidate` baseline. It replaces the first rendered prompt package with one that removes all recorded agent and tool events. Baseline traces are reused without rerunning inference. A manifest frozen before the intervention assigns 46 pairs to a byte-identical-prompt control, 17 to the stale-observation primary subgroup, and two to an unobserved-irrelevant-history subgroup.

All six intervention files have exact W0/W1 coverage and zero call or parse errors. Paired effects use 1,000 deterministic bootstrap resamples over model-qualified source-task clusters. On the 17 stale-observation pairs, task-only minus serialized-candidate changes NSR by +.412 [.118, .706], PCR by -.412 [-.706, -.118], SFC by +.118 [-.118, .412], and query cost by +.588 [.176, 1.000]. On the 46 controls, all 92 first raw outputs are byte-identical; 86/92 complete behavior signatures agree, and aggregate SFC and query cost are unchanged. This control verifies first-turn prompt identity, not deterministic equality of every later continuation. Because this package replacement removes the full agent/tool history, it serves as a pilot rather than the paper's primary prompt-package estimand.

### A.3 Dependency-Turn Prompt-Package Replacement Factorial

Factorial v2 uses all 65 model-owned forks and crosses three binary factors: serialized versus reconstructed source roles, disclosed versus hidden candidate commitment, and full versus `stale_removed` rendered prompt packages. The replacement package deletes only the internal model-call record, visible tool call, and visible tool observation for reads that expose the W1-mutated dependency. It retains unrelated source history and never removes the commitment boundary. Nineteen forks contain one or more removable stale-evidence turns; the remaining 46 are prompt-identity controls. Three serving repeats per cell, fork-blocked condition interleaving, and randomized world order produce 1,560 pair trials and 3,120 continuations. The estimand is the total rendered prompt-package replacement effect: deleting these turns also changes prompt length, token position, recency, salience, and local discourse, so the factorial does not separately identify semantic reliance on stale evidence.

| Contrast on 19 stale targets | NSR [95% interval] | PCR [95% interval] | SFC [95% interval] | JointATIS [95% interval] |
|---|---|---|---|---|
| Dependency-turn-deleted package minus full package | +.656 [.500, .738] | -.656 [-.748, -.500] | +.145 [-.032, .281] | +.140 [-.042, .273] |
| History x native roles | +.101 [-.181, .289] | -.101 [-.329, .208] | +.044 [-.111, .161] | +.035 [-.108, .158] |
| History x candidate disclosed | -.320 [-.583, -.071] | +.320 [.132, .564] | -.132 [-.382, .117] | -.140 [-.417, .115] |

The frozen primary population is the 19 stale-evidence targets, the primary contrast is the dependency-turn-deleted rendered package minus the full rendered package, and the primary metrics are NSR and PCR. SFC, query cost, and pair-level separator-read fraction are secondary; interactions and model subgroups are exploratory, and intervals are not multiplicity-adjusted. The primary package-replacement effect is pair-weighted over role and disclosure after averaging repeats within each model-owned fork. Query cost changes by +.754 [.503, .975]. The factorial's mean pair-level separator-read fraction changes by +.570 [.439, .659]; it assigns zero to no-read pairs and is distinct from the pooled-read SQP reported in the main sensing-profile table. The NSR/PCR effects are +.474/-.474 for serialized-disclosed, +.737/-.737 for serialized-hidden, +.518/-.518 for native-disclosed, and +.895/-.895 for native-hidden replay. All six model point estimates also have positive NSR and negative PCR changes. The SFC interval crosses zero. The package-by-role interactions are unresolved, while candidate disclosure attenuates the NSR/PCR package-replacement effect relative to hidden-candidate replay.

All 46 identity-control prompt hashes match between history levels in every role/disclosure setting. Aggregate first-response agreement is 98.9–100% and behavior-signature agreement is 98.9–100% across those settings. Across serving repeats, first-response agreement is 98.5–100% and behavior agreement is 97.4–99.5%. Repeats measure endpoint stability rather than independent model sampling. The primary bootstrap treats the six model snapshots as fixed and uses 1,000 resamples of nine raw source-task clusters after averaging repeats within fork; all selected model/fork observations belonging to a sampled task move together. The point estimate remains pair-weighted rather than equal-model weighted.

Factorial v1 is excluded in full. Its native renderer appended a continuation cue as an adjacent second user turn, causing a protocol failure for Mistral and making the presentation invalid for a strict role-alternating comparison. V2 merges the cue into the trailing user turn, validates strict role alternation for every frozen prompt, and reuses none of the v1 model outputs. The frozen v2 manifest hash, source/result hashes, inference commit, exact analysis-script SHA-256, and `raw_results_modified=false` flag are recorded in the factorial provenance. The inference environment did not uniformly record checkpoint-tree, tokenizer, vLLM-package, container, or CUDA hashes, so the artifact supports exact offline re-analysis but not byte-identical model serving reconstruction.

### A.4 Matched Stale-Semantics and Costly-Action Diagnostic

The oral-upgrade diagnostic freezes 48 model-owned stale-evidence forks across the six local model snapshots, with eight forks per model, 19 raw retail or service-operations task clusters, and three serving repeats. Both conditions use reconstructed native observable roles, hide the candidate commitment, interleave conditions within fork, and randomize W0/W1 order. `oral_full_stale` retains the dependency-bearing observation. `oral_matched_placebo` changes only the key and value exposing the W1-mutated dependency. It preserves event types, tool-call sequence, turn

sequence, rendered role sequence, rendered prompt byte count, and a tokenizer-agnostic lexical-token proxy. It does not guarantee equality under each model's tokenizer or behavioral inertness.

| Population | Metric | Full stale | Matched placebo | Delta [95% interval] |
|---|---|---|---|---|
| All 48 | NSR | .000 | .021 | +.021 [.000, .065] |
| All 48 | PCR | 1.000 | .979 | -.021 [-.065, .000] |
| All 48 | SFC | .000 | .000 | .000 [.000, .000] |
| All 48 | Query cost | .167 | .083 | -.083 [-.220, .027] |
| Costly 15 | NSR | .000 | .000 | .000 [.000, .000] |
| Costly 15 | PCR | 1.000 | 1.000 | .000 [.000, .000] |
| Costly 15 | SFC | .000 | .000 | .000 [.000, .000] |

Effects are matched placebo minus full stale after repeat averaging. The six model snapshots are treated as fixed, and 2,000 deterministic resamples cluster by raw `domain:task_id`. One Meta-Llama retail fork accounts for the aggregate NSR/PCR change; five models have zero point changes. In that fork, the model reads the current order status in the placebo condition and still attempts cancellation in W1, so SFC remains zero. These results do not meet a positive semantic-effect gate. They leave intact the earlier total rendered prompt-package replacement estimand while showing that it cannot be attributed specifically to dependency semantics without additional assumptions.

Fifteen selected units, all from the synthetic service-operations environment and spanning eight raw-task clusters, make the authored-policy-invalid W1 candidate simulator-accepted and benchmark-state-changing rather than state-preservingly rejected. Across two conditions, three repeats, and W1, all 90 continuations invoke a designated write that semantically matches the original contracted candidate under the frozen `semantic_commitment_v4` scorer; 69/90 preserve the exact candidate arguments, while the other 21 differ only in scorer-excluded free-text `reason` or `resolution` fields. All 90 simulator calls return `ok=true`, all 90 satisfy the authored contract's existing `harmful_invalid_write` diagnostic, and zero obtain a separator before commitment. This is a legacy operational contract label rather than an independently validated real-world harm judgment. The costly subgroup rules out state-preserving simulator rejection as the explanation for those trace outcomes. It remains a selected synthetic subgroup and does not estimate external effects or real-world harm prevalence.

The outcome-blind audit export contains all 48 units and all 15 costly actions, but its label fields remain null. It is an audit-ready artifact, not completed independent validation.

### A.5 Recorded-vs-Matched-Synthetic Prefix Provenance Diagnostic

The provenance diagnostic freezes 40 public $\tau^2$ retail stale-observation forks across three fixed local model snapshots (Qwen3-8B, Meta-Llama-3.1-8B-Instruct, Qwen2.5-14B-Instruct), one serving repeat, and serialized-history rendering. Both conditions hold structure, roles, timing, candidate disclosure, the continuation cue, and every tool-call/tool-observation event, including the stale dependency read, byte-for-byte identical. `recorded_serialized` retains the authentic recorded dialogue; `matched_synthetic_serialized` replaces only `agent_message`/`user_message` content with deterministic canonical filler matched to the renderer's exact wire-format byte length. `thinking_mode` was set to `disabled` for all three models as an outcome-blind amendment made before any estimand was observed, correcting an inference-setting omission on a discarded first attempt; this is a no-op for the two non-reasoning models.

| Population | Metric | Synthetic | Recorded | Delta [95% interval] | Verdict |
|---|---|---|---|---|---|
| Aggregate (120 pairs) | NSR | .221 | .042 | -.179 [-.233, -.125] | difference_detected |
| Aggregate (120 pairs) | PCR | .671 | .825 | +.154 [.083, .221] | difference_detected |
| Aggregate (120 pairs) | fork_consistency | .058 | .017 | -.042 [-.092, +.008] | matched_synthetic_equivalent |
| Aggregate (120 pairs) | query_cost | .446 | .058 | -.388 [-.492, -.283] | matched_synthetic_equivalent |

Effects are recorded minus synthetic. The three model snapshots are treated as fixed, and 2,000 clustered resamples (seed 1729) resample the 40 `source_task` clusters. The aggregate NSR and PCR

| Model (40 pairs) | Metric | Synthetic | Recorded | Delta [95% interval] |
|---|---|---|---|---|
| Meta-Llama-3.1-8B-Instruct | NSR | .638 | .125 | -.513 [-.663, -.363] |
| Meta-Llama-3.1-8B-Instruct | PCR | .125 | .500 | +.375 [.200, .550] |
| Qwen2.5-14B-Instruct | NSR | .025 | .000 | -.025 [-.075, .000] |
| Qwen2.5-14B-Instruct | PCR | .888 | .975 | +.088 [-.013, .200] |
| Qwen3-8B | NSR | .000 | .000 | .000 [.000, .000] |
| Qwen3-8B | PCR | 1.000 | 1.000 | .000 [.000, .000] |

deltas both point in the preregistered H1 direction and are statistically resolved: both intervals exclude zero and both exceed the preregistered $\pm.15$ equivalence margin, PCR by only .004. This does not meet the preregistration's own gate for trajectories adding measurement value, which additionally requires model-level directions not attributable to a single model. Meta-Llama-3.1-8B-Instruct alone accounts for essentially all of the aggregate's statistical resolution, with NSR -.513 [-.663, -.363] and PCR +.375 [.200, .550]. Qwen2.5-14B-Instruct shows the same-direction NSR trend, but its interval includes zero and sits within the margin, so it reads as equivalent; its PCR interval also includes zero and is inconclusive rather than resolved. Qwen3-8B shows exactly zero delta on every metric, bit-identical behavior regardless of prefix provenance. This is a known, independently reproduced floor rather than a new anomaly. The same model under the same thinking-disabled setting and the same disclosed-candidate convention in the already-completed $\tau^2$ external-transfer study shows NSR .000, PCR 1.000, and query cost .000, while the identical model under an undisclosed-candidate framing in that same study senses normally, with NSR .30, PCR .6125, and query cost 1.275. The floor tracks the disclosed-candidate prompt convention shared by both designs, not prefix provenance or thinking mode. Fork consistency and query cost both land within their preregistered margins and read as matched-synthetic-equivalent. These results do not meet the preregistration's cross-model generality gate: the resolved aggregate effect is carried by one of three models, with a second unresolved in the same direction and a third contributing no directional information.

### A.6 P1: Semantic Key–Value Factorial v2

This experiment is distinct from the dependency-turn prompt-package replacement factorial above. P1 v1 stopped closed when its mandatory tokenizer materialization rule failed, before an endpoint started or any model request, continuation, analyzer output, or behavioral result existed. V2 was an explicit pre-inference amendment, not a retrospective reinterpretation of v1. It replaced the requirement that all four key-by-value cells have equal token counts with a pairwise gate: at each fixed evidence value, relevant and placebo prompts must have equal total and per-message token counts, a single quoted-key substitution, and identical token IDs and positions outside equal-size key edit windows; within each key arm, supporting and blocking prompts must have a single quoted-value substitution and identical exterior tokens up to the same position shift. The supporting-to-blocking token and edit-window deltas must also match across key arms.

The `status` candidate order remained unchanged. For `severity`, 12 candidates were appended after the 11 v1 candidates when two separately prompted, mutually label-blind Codex evaluations both accepted them under an authored neutrality rubric. Raw agreement was 19/30 (63.3%), including 12 joint accepts. These are AI-agent judgments, not human or domain-expert validation. Candidate generation was knowingly informed by the v1 tokenizer failure and favored multi-segment identifiers; all 12 retained additions were locally known to pass the one unmatched Mistral fork before the v2 package was committed. The cluster qualification was therefore a full environment-bound confirmation, not a first-look feasibility test. It evaluated all 361 frozen candidate–fork combinations and 2,888 rendered prompt cells without inference, then selected the first passing key mechanically: `source` for all 14 retail `status` forks, `settings` for eight service-operations `severity` forks, and `set_name` for the remaining Mistral fork.

The frozen behavioral panel contains six fixed model snapshots, 23 model-owned forks, and nine raw `domain:task` clusters. Two scaffolds, two key levels, and two value levels give 184 fork–condition cells. Three serving repeats yield 552 W0/W1 pair trials and 1,104 continuations. Validation found complete coverage and zero protocol, call, parse, or HTTP-retry errors. The analysis first averages repeats within each model-owned fork and condition. It gives the two domains equal

weight, models equal weight within domain, raw tasks equal weight within model–domain, and scaffolds equal weight. Percentile intervals use 10,000 seed-1729 Bayesian bootstrap draws over raw `domain:task` clusters while treating the model panel as fixed.

The three preregistered estimands are $\delta_{NSR} = NSR(P,S) - NSR(R,S)$, $\delta_{PCR} = PCR(R,S) - PCR(P,S)$, and $\gamma_{PCR} = [PCR(R,S) - PCR(R,B)] - [PCR(P,S) - PCR(P,B)]$, where $R/P$ denote relevant/placebo keys and $S/B$ supporting/blocking values. Semantic GO required all three aggregate estimates to be at least +.15 with 95% lower bounds above zero, both domain-specific direct contrasts to be at least +.10, no primary sign reversal in leave-one-task or leave-one-model analysis, and at least four discordant raw-task clusters for every estimand. Underidentification takes precedence over a practical-null interpretation.

Table 2: Confirmatory P1 semantic-factorial estimands after repeat averaging. The full conjunctive decision is `UNDERIDENTIFIED`; `SEMANTIC_GO=false`.

| Estimand | Estimate | 90% interval | 95% interval | Discordant clusters | Preregistered status |
|---|---|---|---|---|---|
| delta NSR | .006944 | [-.005843, .025720] | [-.006432, .028300] | 2/9 | underidentified |
| delta PCR | .006944 | [-.005912, .025344] | [-.006418, .028303] | 2/9 | underidentified |
| gamma PCR | .232639 | [.174636, .293334] | [.169775, .300683] | 4/9 | not flagged underidentified |

Both direct 90% intervals lie within the prespecified $\pm.10$ practical-null margin, but seven of nine clusters have exactly zero direct effect and one cluster points in each direction. The preregistered sparse-discordance rule therefore prevents calling this a demonstrated null. The positive aggregate PCR interaction also does not satisfy the conjunctive gate. It is concentrated in the retail `status/source` edit package: retail $\gamma_{PCR} = .4792$ [.3535, .6155] has three discordant clusters and is underidentified, while service operations is -.0139 with one discordant cluster and is underidentified. Their service-operations-minus-retail package difference is -.4931 [-.6283, -.3677], has four discordant clusters, and is not flagged underidentified. All retail forks use `status: pending` to `shipped` and `cancel_pending_order`; all service-operations forks use `severity:SEV2/SEV3` to `SEV1` and `resolve_mitigated_incident`, with `settings` or `set_name` as placebo. Domain, dependency field, value family, action contract, and placebo family are consequently bundled. The contrast is edit-package-associated heterogeneity, not an identified pure domain effect.

Table 3: P1 domain/edit-package and scaffold stratification. Scaffold interactions lie within the +/-.15 margin, but sparse discordance and the large domain/edit-package interaction prevent a preregistered invariance claim (`approximately_invariant=false`).

| Scope | delta NSR | delta PCR | gamma PCR | Interpretation |
|---|---|---|---|---|
| Retail package | .0278 | .0278 | .4792 | three discordant clusters; underidentified |
| Service-operations package | -.0139 | -.0139 | -.0139 | one discordant cluster; underidentified |
| Service operations minus retail | -.0417 | -.0417 | -.4931 | four discordant clusters; not flagged underidentified |
| Serialized/disclosed scaffold | -.0139 | -.0139 | .2153 | stratified estimates underidentified |
| Native/hidden scaffold | .0278 | .0278 | .2500 | stratified estimates underidentified |
| Native/hidden minus serialized/disclosed | .0417 | .0417 | .0347 | gamma 90% interval [.0154, .0669] |

Sensitivity checks preserve the interpretation rather than upgrading it. Removing `retail:t004` reverses both direct signs and lowers $\gamma_{PCR}$ to .1097; removing Meta-Llama also reverses both direct signs. Pair weighting gives direct estimates of .0181 whose 95% intervals cross zero and $\gamma_{PCR} = .1920$ [.0359, .3916]. Relevant-supporting minus placebo-supporting SFC and JointATIS are each -.0139, each has only one discordant cluster, and both are underidentified. Thus `downstream_reliability_language_permitted=false`. A positive PCR interaction can arise because blocking evidence suppresses commitment without producing separator acquisition; it is not by itself evidence of correct sensing, strict reliability, safety, or an internal binding mechanism.

The inference freeze is commit `01fdbead5bb5a9d96d1b05acef09cf56cdc6dd7e`, the separate authorization is `c4faf8bb3eb35575a085f6d33ea71b87bcb66fcf`, and the evidence commit is `9b3d1b857f261fb44378e442db0d482a686094e4`. The report is under `reports/p1_semantic_factorial_v2/cluster_evidence/p1_semantic_factorial_v2_c4faf8b_20260803T164629Z/run_`

dir/report/; primary_estimands.csv, stratified_estimands.csv, leave_one_out.csv, pair_weighted_sensitivity.csv, secondary_reliability.csv, validation.json, and provenance.json support the statements above. The 872,050-byte evidence capsule has SHA-256 56b148007db93dc1bfafbe3623a68a659432803d9a1dac2f0cdf302caec708fa. Per-model result seals, manifest and prompt/token hashes, model-tree and tokenizer identities, and endpoint/environment records are preserved; the records identify an offline H100 environment with vllm 0.12.0 and Transformers 4.57.6. This is stronger inference provenance than the earlier diagnostics, but it still does not supply an independently portable container or human validation of the placebo semantics.

### A.7 Trajectory-Value Robustness

The preliminary full-history contrast contains 17 model-fork pairs but only seven raw task IDs, six models, and 16 raw fork IDs. Its NSR/PCR direction survives raw-task clustering (+.412/-.412), model clustering (+.412/-.412), equal-model weighting (+.458/-.458), identity-control drift adjustment (+.412/-.412), and the current matcher's 19-pair sensitivity (+.421/-.421); the corresponding reported intervals exclude zero. Leave-one-task and leave-one-model point estimates also preserve direction. These are re-analyses of one selected sample, not independent replications.

The factorial supersedes the 19-pair sensitivity as the primary prompt-package analysis because it freezes those 19 targets before inference, replaces the rendered package by deleting dependency-bearing stale read turns, separates role format from candidate disclosure, and measures serving stability. It does not solve the sample-size boundary: the target set remains small, model-owned, and selected from successful commitments. The replacement also changes prompt length, token position, recency, salience, and discourse. NSR and PCR are often coupled rather than independent confirmations, and neither the preliminary nor factorial analysis establishes a nonzero SFC effect or an internal model mechanism.

### A.8 External $\tau^2$ Trajectory Transfer

The external panel begins from one pinned public $\tau^2$-Bench retail submission (Barres et al., 2025) generated by a Claude Sonnet 4.5 source agent and GPT-5.2 user simulator. From the eligible successful trajectories, deterministic seed 1729 selects eight source tasks each for cancellation, address modification, item modification, return, and exchange commitments, yielding 40 stale-observation forks. This balanced challenge panel does not represent the source submission's natural commitment prevalence. The source file, normalized trajectories, contracts, fork panel, and manifests are stored with hashes and require no runtime network access.

Three context-compatible local snapshots—Qwen3-8B, Meta-Llama-3.1-8B-Instruct, and Qwen2.5-14B-Instruct—evaluate every fork. The final v2 panel replaces a context-incompatible Llama-3 snapshot before inference; the panel and all other settings are unchanged, and no v1 result is reused. serialized_candidate retains all user-visible source trajectory events. user_only_candidate retains all source user turns and the identical unexecuted-candidate scaffold while removing source agent and tool events. All six model-condition files contain 80 continuations with complete W0/W1 coverage and zero call or parse errors, for 480 rows.

| Metric | Serialized history | User-only history | Delta [95% interval] |
|---|---|---|---|
| NSR | .042 | .350 | +.308 [.217, .400] |
| PCR | .833 | .542 | -.292 [-.383, -.204] |
| SQP | .714 | .282 | -.432 [-.737, -.089] |
| Query cost | .058 | 1.271 | +1.213 [.929, 1.550] |
| SFC | .025 | .042 | +.017 [-.025, .067] |
| JointATIS | .017 | .025 | +.008 [-.025, .042] |

Point estimates pool 120 model-fork pairs. The three model snapshots are fixed; 2,000 deterministic bootstrap resamples cluster by the 40 original source task IDs and keep each task's model observations together. NSR increases and PCR decreases for every model point estimate. The aggregate SQP decrease shows that added reading is less selective under the pooled-read definition, and the

SFC interval crosses zero. The same directional rendered prompt-package contrast is therefore observed on this selected source benchmark and scaffold without establishing an authored-policy reliability gain.

The comparison remains coarse. The user-only package deletes relevant and irrelevant source history together and changes prompt length, recency, and action-continuation cues. Neither this comparison nor the dependency-turn package replacement isolates stale semantics. Because all source tasks come from one public submission, all target forks remain retail, and only three evaluation snapshots are included, this pilot is neither cross-domain evidence nor a population estimate over external trajectories.

### A.9 Construction-Source Diagnostic

The matched construction-source audit fixes all 116 commitments in the frozen stored-success generator pool and the exact two contract mutations per commitment, then varies the information supplied to construction. Every selected trajectory is also successful under the current semantic scorer. Two Qwen3 trajectories (`t005` and `t008`) are semantic successes but have frozen stored failure labels, so they remain outside this matched generator pool; the six-model semantic static-success aggregate therefore contains 118 successes. Recorded prefixes, task plus recorded candidate, minimal synthesized prefixes, and contract-only oracle-instantiated candidates each produce 232/232 valid, authored-strict-policy-divergent, separable, oracle-checkable risky forks, and validation confirms identical mutation-signature sets across variants. All 65 behavioral replay pairs join exact source commitment and mutation signatures; their frozen fork-type labels differ from the current matcher only for Llama `t011` and `t012`, as disclosed above. The recorded prefix therefore has no structural constructability-yield advantage in this retail pilot. Among the four implemented representations, it preserves the actual events showing whether an authoritative dependency observation appeared before commitment; W1 then makes that represented observation stale (balanced accuracy 1.00 versus .50 for constant-label alternatives). A richer summary could encode equivalent history. This is an event-history label, not a claim about hidden model state. The contract-only condition receives seed-task action arguments and is an oracle upper bound rather than an automatic commitment-discovery method.

## B Generator Comparison

The generator diagnostic uses the same 116-commitment frozen stored-success pool described above. All generators receive the same immutable trajectory pool and use the same prefix-equivalence, action-divergence, separator, and oracle checks. Risky attempts and controls are reported separately.

| Generator | Commitments | Risky attempts | Valid risky | Risky valid rate | Controls | Invalidation types | Signatures |
|---|---|---|---|---|---|---|---|
| Contract-guided | 116 | 232 | 232 | 1.000 | 123 | 6 | 14 |
| Random mutation | 116 | 116 | 69 | .595 | 0 | 4 | 10 |
| Generic TOCTOU | 116 | 116 | 116 | 1.000 | 0 | 1 | 7 |

Random mutation often changes a fact that is not both action-divergent and separable. Generic TOCTOU matches the contract-guided risky valid rate in this status-heavy domain, so the evidence does not support a blanket valid-yield advantage over that baseline. Contract guidance instead enumerates two dependency-specific risky mutations per source commitment where available and adds controls, yielding broader invalidation and signature coverage. Because candidate budgets differ, raw fork counts are not treated as an efficiency comparison.

## C Harness Diagnostics

The wrapper study uses earlier model-conditioned fork sets and matched fork IDs within each model. Its absolute rates should not be compared directly with the shared-panel main table.

For Qwen2.5-14B, the dependency-aware condition improves 12 of 42 risky pairs and the staleness-aware condition improves 9, with no observed pair regressions. Meta-Llama is unchanged under

| Model | Condition | NSR | Query cost | SFC | Paired SFC delta [95% CI] |
|---|---|---|---|---|---|
| Qwen2.5-14B | neutral | .000 | .000 | .000 | – |
| Qwen2.5-14B | dependency-aware | .905 | 1.202 | .286 | +.286 [.143, .429] |
| Qwen2.5-14B | staleness-aware | 1.000 | 1.310 | .214 | +.214 [.095, .357] |
| Qwen2.5-14B | oracle diagnostic | .905 | 1.155 | .476 | +.476 [.333, .619] |
| Meta-Llama-3.1-8B | neutral | .895 | 2.197 | .079 | – |
| Meta-Llama-3.1-8B | dependency-aware | .921 | 2.289 | .079 | .000 [.000, .000] |
| Meta-Llama-3.1-8B | staleness-aware | .974 | 2.368 | .079 | .000 [.000, .000] |
| Meta-Llama-3.1-8B | oracle diagnostic | .921 | 2.276 | .079 | .000 [.000, .000] |
| Llama-3-8B | neutral | .675 | 1.188 | .075 | – |
| Llama-3-8B | dependency-aware | .750 | 1.375 | .075 | .000 [.000, .000] |
| Llama-3-8B | staleness-aware | 1.000 | 1.838 | .050 | -.025 [-.100, .050] |
| Llama-3-8B | oracle diagnostic | .800 | 1.400 | .025 | -.050 [-.125, .000] |

all three interventions. Llama-3 is unchanged under dependency-aware sensing; staleness-aware sensing improves one pair and worsens two, while its oracle condition worsens two. These patterns motivate profile-level bottleneck interpretation rather than a universal wrapper claim.

Selectivity remains limited. For Qwen2.5-14B, dependency-aware interception occurs on 38 of 50 control continuations and staleness-aware interception on all 50; the oracle intercepts none. The oracle uses fork metadata and is neither deployable nor a performance ceiling.

## D Prompt and Output-Protocol Failures

Mistral-7B and Qwen2.5-7B each cover all 72 explicit-caution units structurally, but their result files contain seven and one JSON parse errors, respectively. Mistral emits multiple concatenated tool-call objects on task `t015`; Qwen2.5-7B emits one malformed object on stale task `t021`. Their neutral runs are error-free under the same model-specific protocol.

In the service-operations instantiation, Mistral produces four parse errors on two fork IDs, in both W0 and W1, after emitting multiple objects. All affected requests return HTTP 200. We preserve the failed files, do not modify the parser, and exclude those model-condition blocks from clean aggregates.

## E Audit Scope

The legacy single-user-accepted retail construction review has 72 label rows from three earlier model-conditioned exports: 48 risky rows and 24 no-risk-control rows. Repeated IDs reduce this to 27 unique forks. Twenty-two occur in the primary shared panel, covering 15 initial-dependency and 7 no-risk units, with no reviewed overlap for stale or irrelevant-change units. Forty-nine sampled coarse replay judgments were accepted. Fine-grained failure-reason completeness was accepted for 27 of 49 sampled failures. This author-accepted material is not independent human validation.

The service-operations audit packet contains all 56 panel units in deterministic randomized order and omits model outcomes. A context-isolated Codex subagent restricted to the packet, contracts, schema, manifest, and validator accepted 56/56 constructions and recorded no issue codes. This is model-assisted review, not independent human annotation or an agreement estimate.

The primary retail panel also has an earlier full outcome-blind model-assisted semantic audit under a different rubric. Its row-level report covers all 72 unique raw units and records 21 `accept`, 43 `minor_issue`, and 8 `major_issue` verdicts. All 72 pass environment-relative dependency completeness, prefix equivalence, and W0/W1 validity. The eight major verdicts identify the same legacy extra separator candidates removed by the v3 scoring normalization. The minor findings record the authored no-write policy versus state-preserving rejected-invocation policy choice, 14 risk-severity label conflicts, and smaller metadata or description inconsistencies. The audit is not human validation and does not estimate inter-annotator agreement.

Two separately instantiated, context-isolated Codex agents independently completed the stricter `p0_contract_audit_v1` packet. Overall labels match on 60/72 units (83.3%; Cohen's $\kappa =$

.443): one annotator assigns 61 major and 11 minor verdicts, while the other assigns 59 major, 3 minor, and 10 valid verdicts. The 70 field-level disagreements span 44 units and concentrate in dependency completeness (39 disagreements; 33/72 exact agreement; $\kappa = .278$). W0 acceptability agrees on 60/72 ($\kappa = .665$), and W1 acceptability agrees on 68/72 ($\kappa = .806$).

The annotators agree on every mutation-validity, separator-sufficiency, and state-preserving-rejection/no-write-policy field. Both judge all 72 mutations valid. Both also mark the policy field major for all 48 risky units and valid for all 24 controls, chiefly because the authored strict W1 behavior set excludes state-preserving rejected invocations. Thus the dominant finding is a construct-policy objection rather than an invalid-state-mutation finding.

A third distinct, context-isolated Codex agent adjudicated all 72 rows only after both label files and their agreement report were frozen. It resolves every one of the 70 disputed fields without overriding an agreed field. Final overall decisions are 61 `major_issue`, 11 `minor_issue`, zero `valid`, and zero `uncertain`; all 48 risky units are major, while controls contain 13 major and 11 minor. Mutation validity and authority-source validity are 72/72 valid; dependency completeness is 24 major and 48 minor. No risky unit enters either the valid-only or valid-plus-minor inclusion set. We therefore report audit-filtered SFC and JointATIS as not estimable, not zero; the frozen 48-pair values remain authored-policy sensitivities alongside the state-preserving-rejection sensitivity. These are judgments from three Codex agents in a shared model family, not human or domain-expert validation, and $\kappa$ does not estimate inter-human reliability. A stricter blinded protocol is released as future validation infrastructure, uses author-independent entity-closure evidence and a staged freeze gate, and contributes no eligible-person annotations to the reported results.

### E.1 Fail-Closed Oracle Sensitivity: State-Preserving Rejection

| Model | Strict SFC | Verified matching W1 rejection | State-preserving sensitivity | JointATIS | Rejection-joint sensitivity |
|---|---|---|---|---|---|
| Qwen3-8B | .125 | 31/48 | .771 | .125 | .125 |
| Meta-Llama-3.1-8B | .104 | 4/48 | .188 | .104 | .104 |
| Mistral-7B | .063 | 12/48 | .313 | .063 | .063 |
| Llama-3-8B | .083 | 4/48 | .167 | .083 | .083 |
| Qwen2.5-7B | .292 | 9/48 | .479 | .271 | .271 |
| Qwen2.5-14B | .042 | 43/48 | .938 | .042 | .042 |

The trace-verified sensitivity retains the strict W0 requirement. W1 additionally receives mechanical credit only when the first parsed designated-write invocation occurs before a separator, semantically matches the candidate, returns the expected simulator rejection, and leaves the benchmark application-state snapshot unchanged. Each credited trace is reconstructed from its W1 fork and exact arguments five times; every response, state delta, and frozen final state must match. This tightening rejects six legacy `failed_write_probe` flags—four Mistral calls with empty arguments and two Llama returns missing candidate item arguments—although those pairs already fail W0, so the historical pair-level values happen to remain unchanged.

All 48 canonical W1 candidate invocations are rejected with `precondition_failed`, no exception, and an empty benchmark application-state delta. Three W0 canonical candidates also succeed with no application-state delta, preserving separate task-semantic concerns from the audit. The sensitivity is **not** SFC, ATIS, terminal success, correctness, utility, or safety: the simulator records calls, unmodeled effects may remain, and replay stops before recovery. Credit changes no JointATIS value because every newly credited invocation precedes a separator. In the enriched costly-action subset, 15 model-owned forks from eight raw-task clusters, all from synthetic service operations, yield 90 repeated pair trials across two conditions and three repeats; every W1 first invocation is simulator-accepted and benchmark-state-changing, and all four authored-policy/sensitivity/joint endpoints are zero.

The offline `p0_contract_policy_multiverse_v1` report closes the mechanical replay surface without adding model calls or human labels. It verifies 63/63 fork witnesses—48 shared-panel retail forks plus 15 costly service-operations forks—and all 378 associated frozen W1 trace witnesses. Each fork record checks the shared-prefix hash, one contract-predicate mutation path, executable W0-pass/W1-fail candidate divergence, a differentiating authoritative separator, candidate response, and benchmark-state delta. In the shared traces, strict SFC is 34/288 and the state-preserving-rejection ceiling is 137/288, while strict-JointATIS and rejection-joint counts are both 33/288 because the commitment boundary does not move. In the costly traces, all 90/90 writes are

accepted, state-changing, and decision-invalid, and both strict SFC and JointATIS are zero. These are deterministic contract/simulator witnesses, not construct validation, terminal utility, safety, or a prevalence estimate.

## F METRIC INTERPRETATION

The first-commitment ungrounded rate uses all valid static trajectories as its denominator. It must not be described as a fraction of static successes. `strict_static_live_gap` is `static_success − strict_fork_consistency`; the paper-facing `fork_consistency` field is a deprecated numerical alias retained for artifact compatibility. `live_fork_failure_rate` is `1 − strict_fork_consistency`. These are different quantities.

Static success is trajectory-weighted, while SFC and JointATIS are risky-pair-weighted. JointATIS is one only when W0 and W1 both obtain a successful separator before the scorer's first parsed invocation of a contract-designated write-capable tool and both satisfy the authored strict behavior policy. It is therefore bounded above by SFC. Observed incomplete, duplicate, invalid-world, evidence-malformed, and protocol-failure pairs score zero and remain auditable in its denominator; expected-panel validation catches completely absent trials. Some tasks contribute several forks. Cluster bootstrap intervals quantify fixed-panel resampling variability and do not form a joint interval for the strict static-live gap.

`stability_cue_read_rate` is the fraction of continuations in the stored `no_risk_control` condition with any pre-commitment read. It is not a pure over-sensing rate: the prompt guarantees no post-prefix dependency change, but 10 of 16 retail control prefixes lack an authoritative observation of at least one action dependency. The legacy artifact fields `no_risk_over_sensing_rate` and `atomic_guard_*` are retained only as compatibility aliases; the latter correspond to the paper-facing trace-verified state-preserving-rejection sensitivity and do not establish atomicity.

## G SEMANTIC SCORING AND IMMUTABLE RE-ANALYSIS

Static scoring protocol `retail_semantic_v2` requires a successful intended tool with task-defining arguments. It ignores free-text `reason`, collapses string whitespace, removes a terminal sentence period from `new_address`, and compares `item_ids` without order. Relative to the frozen legacy labels, this changes Qwen3 tasks `t005` and `t008` from failure to success; no other primary model changes a valid-task static label.

Represented grounding uses dependency-observation matcher `semantic_arg_subset_v2`: a successful authoritative read must match every contract-required argument, while unrelated extra arguments are ignored. The earlier exact-argument matcher produced two false negatives for Llama payment-change trajectories (`t011` and `t012`). Correcting those metadata labels changes Llama grounded static success from 5/22 to 7/22 and its first-commitment ungrounded count from 15/22 to 13/22; model outputs and static outcome labels are unchanged. It also changes one service-operations Meta-Llama grounding count from 1/24 to 2/24.

Replay protocol `semantic_commitment_v4` uses the same task-defining action identity. In W0, the first parsed invocation of a contract-designated write-capable tool must semantically match the source commitment and be simulator-accepted. In W1, any such parsed invocation fails the primary authored oracle; withholding passes. This fail-closed rule defines strict no-write fork consistency, avoids crediting an unchecked alternative write, and does not score explanation quality or post-invocation recovery. A separator counts only when its matching observation returns `ok=true`, contains every required separator argument, and the canonical observation differs between W0 and W1. Accepted extra arguments do not invalidate the match. Failed or non-differentiating reads remain in query cost but not NSR or SQP's separator numerator. NSR is pair-relative and does not certify simultaneous separation of every sibling continuation under a contract.

JointATIS implementation `pair_joint_atis_v1` consumes these frozen scorer outputs rather than re-proving separator executability from summary rows. Its boundary is deliberately conservative: a state-preserving simulator-rejected invocation remains the first parsed invocation of a

contract-designated write-capable tool, so a later observation cannot retroactively count as pre-commitment acquisition. On the primary shared panel JointATIS is 2–13/48 across configurations; it equals SFC for five configurations and is one pair lower for Qwen2.5-7B (13/48 versus 14/48).

The earlier `semantic_commitment_v2` correction changed nine of 432 model-by-panel pair labels by repairing action identity and the invalid-world oracle. Version 3 preserves those behavior labels and normalizes legacy separator declarations: eight of the 48 risky retail units listed one extra contract candidate read whose output did not differ between their worlds, so v3 removes that candidate. Version 4 then replaces exact full-dictionary separator-argument equality with required-argument subset matching. This prevents a successful authoritative read from being discarded merely because its accepted call includes an extra argument. In the primary panel, v4 changes NSR/SQP/PCR labels for six of 576 risky continuations, all from three Llama-3 forks; SFC, query cost, and static labels are unchanged. Original JSONL files remain read-only, and derived files retain legacy labels, scoring versions, source and output hashes, and a no-modification flag.

Service-operations contract metadata originally marked failed precondition writes as side-effectful, although every environment implementation returns before mutation. The metadata is corrected to the legacy fields `safe_to_probe=true` and `side_effect_on_failure=false`, then applied in a deterministic offline re-score. This changes only the legacy safe-probe, unsafe-premature-action, and harmful-write diagnostic fields; NSR, SQP, PCR, query cost, and SFC are unchanged. The correction record retains both contract hashes, affected fields, and confirmation that no model calls were made.

## H Evaluation-Unit Comparison

The closest prior and concurrent evaluations differ primarily in their unit, intervention boundary, and reported endpoint. They cover stale-state execution, pre-execution change, observation validity, and recheck timing (Lilienthal & Hong, 2025; Jiang et al., 2026; Wang et al., 2026; Cheng et al., 2026); evidence grounding (Sheng et al., 2026; Flynt, 2026); authorization and abstention (Santos-Grueiro, 2026; Liu et al., 2026); and procedural or causal trajectory analysis (Keren et al., 2026; He et al., 2025; Shah, 2026). The table states operational distinctions rather than priority claims.

| Evaluation | Unit | Focal boundary | Acquisition endpoint |
|---|---|---|---|
| TOCTOU-Bench | Stale-state task | Outcome | Not focal |
| Atomicity for Agents | Browser trajectory | Pre-execution | Not focal |
| ContractBench | Observation artifact | Temporal validity and integrity | Not separate |
| TicToc | Recheck decision | Elapsed-time/tool-call trade-off | Not pair-relative |
| EnvTrustBench | Executable workspace | Evidence-grounding defect | Not at parsed write |
| GroundEval | Evidence path | Final claim | Path coverage |
| Commit-Time Authorization | Authorization witness | Durable effect | Not focal |
| AgentAbstain | Act–abstain task pair | Outcome | Not focal |
| TASTE | Synthesized tool sequence | Procedural coverage | No |
| TRAJECT-Bench | Synthesized trajectory | Dependency/order coverage | No |
| Causal Agent Replay | Replayed trajectory | Outcome attribution | No |
| This work | Represented-prefix world pair | First parsed write | Yes: NSR/PCR |