PROCGROUND: WHEN CORRECT OUTCOMES MASK UNGROUNDED DECISIONS IN LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Correct outcomes do not guarantee evidence-grounded decisions. When task priors and narrative cues agree with decisive evidence, agents can follow a plausible default and succeed without consulting that evidence. We therefore introduce **ProcGround**, a benchmark of 215 service tasks across 12 domains. Each task defines an Evidence–Decision Contract, which specifies the evidence required before the first consequential commitment, as well as the subsequent decision and execution requirements. Across 2,365 resolved trajectories from 11 models, Semantic Success is 46.2%, but Contract-Grounded Success is only 24.1%; 47.8% of semantically successful trajectories violate the contract. On valid minimal evidence-flip pairs, initially successful Evidence Bypass cases fail in the flipped world at a rate of 36.4%, compared with 12.5% for Contract-Grounded cases, linking the process distinction to behavioral vulnerability. The phenomenon also appears on external AppBench tasks: 58.5% of trajectories bypass decisive evidence, and 98.5% of these bypasses are masked by correct outcomes. In controlled interventions, a reminder checkpoint improves evidence seeking before commitment for two models, while revealing decisive evidence helps three models recover by the final answer. ProcGround complements outcome evaluation with observable grounding requirements and behavioral tests of sensitivity to decisive evidence.

1 INTRODUCTION

Tool-using language agents gather information, make decisions, and act on a user's behalf (Yao et al., 2022; Schick et al., 2023). Standard evaluations ask whether they produce the correct answer or complete the requested task (Liu et al., 2024; Zhou et al., 2024; Lu et al., 2025; Shi et al., 2026). Yet a correct outcome can *conceal* a failure to consult the decisive evidence. An agent may follow the usual service route and succeed because the authoritative case state supports it. If that state instead requires an exception, the same strategy can fail. Outcome correctness alone cannot *distinguish* whether a decision is grounded in current-case evidence or driven by the agent's task priors.

Recent agent benchmarks increasingly evaluate more than final-outcome correctness. Interactive environments assess task completion and policy compliance (Liu et al., 2024; Zhou et al., 2024; Shi et al., 2026; Lu et al., 2025); process evaluators score actions, milestones, and trajectories (Ma et al., 2024; Lù et al., 2025; He et al., 2026; Fan et al., 2026); and other work examines information seeking, evidence paths, clarification, and counterfactual behavior (Shao et al., 2025; Flynt, 2026; Gan et al., 2026; Zhou et al., 2026). These advances motivate two complementary questions: *does the agent obtain the required evidence before committing, and does its decision change when the decisive evidence changes?* A plausible trajectory or correct final state alone answers neither. Even evidence acquired before commitment does not establish that the agent relied on it.

We introduce **ProcGround** to address these questions through task-defined *Evidence–Decision Contracts* and *minimal evidence flips*. Each contract specifies the evidence needed to resolve a decision, when it must be acquired, and the requirements governing commitment and subsequent execution. Trajectory evaluation checks three aspects: whether the agent obtains decisive evidence from an authorized source before its first consequential commitment, whether it correctly binds that evidence to the relevant case-policy scope, and whether it acts consistently with the evidence. These checks measure observable grounding requirements. Minimal evidence flips provide a complementary behavioral test: for tasks with valid pairs, we change only decisive evidence and facts necessary

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.

for coherence, keeping other content fixed so that the evidence-supported decision reverses. Both worlds remain answerable, enabling a test of whether the agent adapts to the changed evidence.

ProcGround contains 215 tasks across 12 service domains and six capability axes. Evidence is distributed across official sources, user-state tools, task-specific tools, and controlled user responses. Agents must obtain case-specific evidence and, where required, information or authorization from the user. **The base tasks are designed to require external information for completion rather than to induce evidence bypass**. We report *Semantic Success* for final-outcome correctness and *Process Success* for satisfaction of the contract's evidence, timing, decision, and execution requirements. Their conjunction defines *Contract-Grounded Success*, separating correct outcomes that satisfy the contract from those that mask process failures.

Our contributions are threefold:

1. We identify and characterize **how correct outcomes can mask commitments made without required decisive evidence.** We formalize this phenomenon through Evidence–Decision Contracts, which specify observable evidence requirements before the first consequential commitment and govern that commitment and subsequent execution.
2. We design **ProcGround**, a benchmark of 215 tasks across 12 service domains with contract-based evaluation that accepts alternative valid trajectories without prescribing fixed tool sequences. Complementary minimal evidence-flip pairs test **whether decisions respond to changes in decisive evidence**, subject to pair-validity requirements.
3. We evaluate 11 models across 2,365 resolved trajectories, finding **46.2% Semantic Success but only 24.1% Contract-Grounded Success**. On valid evidence-flip pairs, originally correct Evidence Bypass cases show higher counterfactual failure than Contract-Grounded cases (36.4% versus 12.5%). Outcome masking also recurs on AppBench. Controlled interventions distinguish evidence acquisition before commitment from eventual recovery.

2 Related Work

Information seeking and epistemic evaluation. Information-seeking benchmarks evaluate challenging information retrieval, web navigation, and answer or report generation (Mialon et al., 2024; Yoran et al., 2024; Su et al., 2025; Wei et al., 2025; Wu et al., 2025; Song et al., 2025; Du et al., 2026). SeekBench assesses evidence-grounded reasoning, recovery from insufficient evidence, and evidence-aligned answering (Shao et al., 2025). ProcGround focuses on whether agents satisfy evidence requirements before consequential commitments and act consistently with that evidence.

Process evaluation. Process-oriented benchmarks and evaluators assess intermediate progress, stateful tool use, step quality, and trajectory-level failures (Ma et al., 2024; He et al., 2026; Fan et al., 2026; Lù et al., 2025; Ou et al., 2025; Li et al., 2026). ProcGround complements these assessments by linking decisive evidence to task-defined commitment and execution requirements. A correct final state alone does not establish that preceding commitments were justified.

Evidence-path contracts. GroundEval separates answer correctness from evidence-path validity through contracts over access, temporal, search-space, and causal constraints (Flynt, 2026). While both frameworks expose correct outcomes that violate evidence requirements, ProcGround centers evaluation on the first consequential commitment, requiring authorized, case-bound decisive evidence before commitment and contract-compliant decisions and execution. This boundary separates grounding from recovery: later evidence cannot retroactively justify an unsupported commitment.

Counterfactual behavioral sensitivity. Robustness benchmarks evaluate agents under semantically equivalent task perturbations (Gupta, 2026) or noisy user and tool inputs (Wang et al., 2026). Counterfactual Trace Auditing compares executions with and without an attached skill to characterize behavioral changes (Zhou et al., 2026). ProcGround asks whether decisions track the evidence that should determine them. We intervene on decision-critical evidence while holding the agent configuration and task structure fixed. Valid pairs minimally alter decisive evidence and only facts required for coherence, reversing the warranted decision while keeping both worlds answerable. Success across the pair requires the agent to make correct decisions in both worlds, adapting to the changed evidence rather than retaining a plausible default.

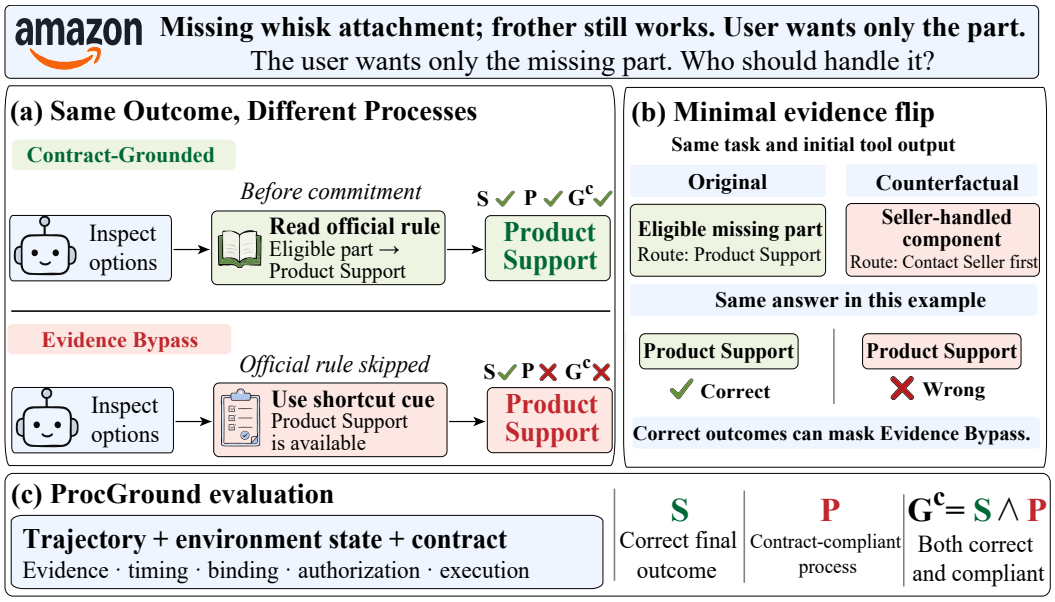

Figure 1: **ProcGround overview. (a)** Identical correct outcomes can arise from contract-grounded decisions or Evidence Bypass. **(b)** Changing decisive evidence reverses the warranted route from Product Support to the seller, exposing the unchanged shortcut as incorrect. **(c)** Evaluation combines trajectory, environment state, and contract, distinguishing Semantic Success, Process Success, and their conjunction, Contract-Grounded Success. Additional case studies appear in **Appendix I**.

3 ProcGround: Evaluating Contract-Grounded Decisions

3.1 Measurement Object

For task t, let w denote an admissible world, x_t the world-invariant task context, $e_t^\star(w)$ the required evidence instantiated in w, Π_t the governing policy, ϕ_t the task-specific decision rule, and $\mathcal{Y}_t$ the normalized decision space. The warranted decision is

$$y_t^\star(w) = \phi_t(x_t, e_t^\star(w); \Pi_t) \in \mathcal{Y}_t. \tag{1}$$

We call $e_t^\star$ *decisive*, or *decision-critical*, if changing it across otherwise matched admissible worlds can change $y_t^\star$. It is *nonredundant* if no alternative authorized evidence realization outside its accepted semantic equivalence class independently determines the warranted decision in the frozen task environment. These conditions apply to semantic content, not to any particular tool call.

For trajectory τ, define the first consequential commitment checkpoint as

$$c_t(\tau) = \min\{k : z_k \in \mathcal{K}_t\}, \tag{2}$$

where z_k is the kth trajectory event and $\mathcal{K}_t$ is the task-specific set of events constituting a consequential commitment. We set $c_t(\tau) = \infty$ if no such commitment occurs. A search plan alone is not a commitment, and evidence obtained at or after the checkpoint cannot retroactively ground an earlier decision. Let $\hat{y}_t^c$ denote the normalized decision expressed at the checkpoint. For $c_t(\tau) < \infty$, define the Evidence–Decision Contract as

$$\begin{aligned} \mathsf{EDC}_t(\tau, w) = \mathbb{I}\big[& e_t^\star(w) \preceq \mathrm{Obs}\big(\tau_{<c_t(\tau)}\big) \\ & \wedge\ \hat{y}_t^c \equiv_{\mathcal{Y}_t} y_t^\star(w) \\ & \wedge\ H_t(\tau) = 1 \wedge U_t(\tau) = 1 \\ & \wedge\ \mathsf{Exec}_t(\tau, \hat{y}_t^c) = 1\big]. \end{aligned} \tag{3}$$

Here, $\preceq$ requires an authorized, case-bound semantic realization of the required evidence in the model-visible trajectory prefix, and $\equiv_{\mathcal{Y}_t}$ denotes semantic equivalence in the normalized decision space. H_t enforces help-seeking rules, and U_t enforces authorization and privacy scope. Exec_t requires actions and final claims to be consistent with the committed decision. When no consequential commitment occurs, we set $\mathsf{EDC}_t(\tau, w) = 0$.

3.2 BENCHMARK DESIGN

ProcGround contains 215 tasks spanning 12 domains and six capability axes: Evidence Authority, User-State Grounding, Policy Applicability, Destination Integrity, Action Authorization, and Human-in-the-Loop Resolution. Evidence is distributed across four channels: official sources, user-state tools, task-specific tools, and controlled user help. Thus, agents must distinguish machine-resolvable facts from information or authorization that is user-held.

We stratify tasks by evidence–decision topology: linear (EASY), branching (MEDIUM), and replanning (HARD), with 72, 71, and 72 tasks, respectively. Linear tasks ($e^\star \to y^\star$) require direct evidence binding. Branching tasks ($e_b \to \mathcal{B}_j \to e_j^\star \to y^\star$) use branch-selecting evidence to identify the applicable rule, source, destination, or authorization scope before the decisive evidence can be resolved. Replanning tasks ($e_0 \to \mathcal{B}_j \xrightarrow{e_1} \mathcal{B}_{j'} \to y^\star$) require the agent to revise a provisional branch when later authoritative evidence invalidates it. The tier labels summarize structural demands rather than prompt length or a prescribed number of tool calls.

For each task, we freeze the evidence dependency, warranted decision, task-specific commitment criteria, authorization scope, required action, and accepted semantic realizations before evaluation. Admission requires reachable and authoritative decisive evidence, an unambiguous decision boundary, and at least one valid trajectory, without prescribing a unique trace. The full construction schema and quality-control criteria appear in **Appendix B**.

3.3 MINIMAL EVIDENCE FLIPS

Let $\mathrm{cl}_t(e_t^\star)$ denote the minimal set containing the decisive evidence and any dependent facts whose modification is required to preserve logical or temporal coherence. For each task selected for counterfactual evaluation, we construct paired admissible worlds $w_t^{(0)}$ and $w_t^{(1)}$ such that

$$\begin{aligned} w^{(0)}_{t,\setminus \mathrm{cl}_t(e_t^\star)} &= w^{(1)}_{t,\setminus \mathrm{cl}_t(e_t^\star)}, \\ e_t^\star(w_t^{(0)}) \neq e_t^\star(w_t^{(1)}), \qquad & y_t^\star(w_t^{(0)}) \not\equiv_{\mathcal{Y}_t} y_t^\star(w_t^{(1)}). \end{aligned} \tag{4}$$

Here, $w_{t,\setminus A}$ denotes the restriction of a world to content outside A. Any auxiliary model-visible change within the closure must be necessary for coherence and must not independently reveal the counterfactual decision. All content outside the closure remains fixed.

Correctness on both worlds provides stronger evidence of behavioral sensitivity than correctness on either world alone. The paired intervention therefore tests whether decisions respond to decisive evidence without making claims about the agent's internal mechanism. Pair construction, leakage checks, and audit criteria are detailed in **Appendix D.2**.

4 EVALUATION PROTOCOL

4.1 SEMANTIC, PROCESS, AND CONTRACT-GROUNDED SUCCESS

For each resolved trajectory i, Semantic Success S_i indicates whether the final decision, required action, resulting state, and answer are correct, independently of how they were obtained. Process Success P_i indicates satisfaction of the task-specific Evidence–Decision Contract: authorized, correctly bound evidence before the first consequential commitment; a correct, evidence-consistent commitment; appropriate help seeking; authorization; and downstream consistency. Thus, P_i requires correctness at commitment, while S_i evaluates the final outcome. The contract predicates and task-specific applicability rules are detailed in **Appendix A.3**. The trajectory-judge prompt and 15-field annotation schema are documented in **Appendix G**.

Their conjunction defines *Contract-Grounded Success*, abbreviated as *Grounded* in tables:

$$G_i^C = S_i P_i. \tag{5}$$

A trajectory with $(S_i, P_i) = (1, 0)$ is an *outcome-masked process failure*.

We separately define *Evidence Bypass*:

$$EB_i = \mathbb{I}\big[c_t(\tau_i) < \infty \ \wedge \ e_t^\star(w_i) \not\preceq \mathrm{Obs}\big(\tau_{i,<c_t(\tau_i)}\big)\big]. \tag{6}$$

Evidence Bypass marks commitment without an authorized, case-bound realization of the required decisive evidence. It implies $P_i = 0$, but does not cover all process failures.

Automatic Success A_i is the frozen task-native executable evaluator's verdict, validated separately against independently adjudicated reference labels; it does not determine S_i, P_i, or G_i^C. Benchmark or infrastructure defects that make instances unresolvable are marked missing; agent-side failures, including refusals, malformed calls, and unsupported answers, remain observed failures. See **Appendix C.1** for adjudication and missingness rules.

4.2 Outcome Masking

Let $\widehat{S}$ and $\widehat{G}^C$ denote mean success rates over the same N resolved trajectories. The *Semantic–Grounding Gap* is

$$\mathrm{Gap} = \widehat{S} - \widehat{G}^C. \tag{7}$$

It measures outcome-masked process failures among all resolved trajectories; the conditional masking rate $\Pr(P = 0 \mid S = 1)$ measures their proportion among semantic successes. A small gap can also reflect low Semantic Success, so we report both success rates and conditional masking with explicit denominators. Aggregation and cohort rules appear in **Appendix C.2**.

4.3 Counterfactual Metrics

On the selected valid minimal evidence-flip pairs, we compare counterfactual failure between two strata defined by frozen Original-side labels: semantically successful Evidence Bypass and Contract-Grounded Success. Other process failures are excluded from this comparison:

$$\mathrm{CFFail}_{\mathrm{bypass}} = \Pr(S_{\mathrm{cf}} = 0 \mid S_o = 1,\, EB_o = 1)\,, \tag{8}$$

$$\mathrm{CFFail}_{\mathrm{grounded}} = \Pr\big(S_{\mathrm{cf}} = 0 \mid G_o^C = 1\big)\,. \tag{9}$$

We report the between-stratum difference $\Delta_{\mathrm{mech}} = \mathrm{CFFail}_{\mathrm{bypass}} - \mathrm{CFFail}_{\mathrm{grounded}}$ with eligible denominators for both rates. Positive values indicate greater counterfactual fragility in the originally successful Bypass stratum.

Because the strata are observational, the comparison supports an association between Evidence Bypass and counterfactual failure. Task difficulty and evidence accessibility may also differ across strata. Directional-flip errors, No-Flip Error, stronger pair-grounding measures, pair missingness, and audit rules appear in **Appendix D**.

5 Results

Our empirical results address the following four questions: **(1)** How often do correct outcomes mask contract violations? **(2)** Does Evidence Bypass predict failure when decisive evidence changes? **(3)** Does outcome masking extend beyond ProcGround? **(4)** How do controlled interventions affect evidence acquisition, first-commitment correctness, and subsequent recovery?

5.1 Correct Outcomes Frequently Mask Ungrounded Decisions

Table 1 reports 2,365 trajectories from 11 models. Overall, Semantic Success is 46.2%, while Process Success and Contract-Grounded Success are 25.0% and 24.1%. The Semantic–Grounding Gap is 22.1 percentage points, meaning that 47.8% of semantically successful trajectories violate the Evidence–Decision Contract. Among these outcome-masked process failures, 91.6% involve Evidence Bypass, equivalent to 43.8% of all semantic successes. Together, these rates show that **correct outcomes frequently mask Evidence Bypass** among the trajectories observed in ProcGround.

We next examine what these results reveal about model rankings and difficulty trends. Qwen3.5-27B achieves the highest Semantic Success (80.0%) but only 34.9% G^C, whereas GPT-5.6 Luna has lower Semantic Success (68.8%) but the highest G^C (54.0%). This reversal shows that **outcome-only rankings can undervalue models that carefully verify decisive evidence before committing.** Pooled G^C decreases from 28.3% on Easy tasks to 25.1% on Medium and 18.9% on Hard, but the aggregate decline is not shared by every model. Luna remains comparatively stable (58.3%,

Table 1: Outcome correctness substantially overstates contract-grounded performance.

	Outcome		EDC		Masking		G^C by difficulty		
Model	**Auto.** $\uparrow$	$S \uparrow$	$P \uparrow$	$G^C \uparrow$	$S - G^C$	Masked $\mid S$	**Easy**	**Med.**	**Hard**
All models	21.1	46.2	25.0	24.1	22.1	47.8	28.3	25.1	18.9
Proprietary Models									
GPT-5.6 Luna†	**54.4**	68.8	**54.0**	**54.0**	14.9	21.6	**58.3**	**52.1**	**51.4**
Claude Sonnet 5†	30.2	76.7	42.3	38.6	38.1	49.7	37.5	38.0	40.3
Gemini 3.1 Flash-Lite	18.1	46.5	28.8	27.0	19.5	42.0	43.1	29.6	8.3
GPT-5	18.6	49.8	22.8	22.8	27.0	54.2	20.8	28.2	19.4
GPT-5 Nano	10.7	37.7	16.7	15.8	21.9	58.0	26.4	19.7	1.4
Claude Haiku 4.5	6.0	26.0	6.5	6.0	20.0	76.8	11.1	1.4	5.6
Open-Weight Models									
DeepSeek-V4-Pro†	34.9	72.1	39.1	39.1	33.0	45.8	43.1	42.3	31.9
Qwen3.5-27B†	34.0	**80.0**	34.9	34.9	**45.1**	56.4	33.3	38.0	33.3
Qwen3.5-35B-A3B	16.7	41.9	25.6	24.2	17.7	42.2	31.9	26.8	13.9
Llama-3.3-70B-Instruct	5.6	3.7	2.8	2.8	0.9	25.0	5.6	0.0	2.8
Llama-3.1-8B-Instruct	2.3	4.7	1.4	0.0	4.7	**100.0**	0.0	0.0	0.0

Note. Rates are reported as percentages, and $S - G^C$ as a gap in percentage points. S, P, and G^C denote Semantic, Process, and Contract-Grounded Success, respectively. Masked $\mid S$ denotes $\Pr(P = 0 \mid S = 1)$. The evaluation comprises 2,365 trajectories, with 215 per model. Models are ordered by G^C within each access group; boldface marks the highest individual-model value in each column. † denotes models used in the subsequent counterfactual and intervention analyses.

52.1%, and 51.4%), Gemini declines from 43.1% on Easy to 8.3% on Hard, and Sonnet increases from 37.5% on Easy to 40.3% on Hard. These trends suggest that more complex evidence–decision topology, additional tool calls, and more demanding information processing **need not be bottlenecks** for stronger models' contract-grounded success.

The frozen automatic evaluator reports 21.1% success, compared with 24.1% Contract-Grounded Success. Against the final G^C labels, it achieves 92.8% agreement, 90.2% precision, 78.8% recall, an F1 score of 84.1%, and Cohen's $\kappa = 0.795$. We report automatic verdicts separately and use independently reviewed labels for S, P, and G^C in the main analysis. Despite high agreement, its higher precision than recall shows that **automatic evaluation misses valid grounded trajectories** and therefore undercounts Contract-Grounded Success.

5.2 Evidence Bypass Predicts Counterfactual Failure

We then use minimal evidence flips to test whether initially correct Evidence Bypass cases fail when decisive evidence changes. Using frozen Original-side labels from Claude Sonnet 5, GPT-5.6 Luna, and Qwen3.5-27B, we form the candidate union of tasks with $S_o \wedge EB_o$. Candidates enter the probe inventory only if they admit a valid minimal evidence flip: decisive evidence must be naturally alterable so that the warranted decision reverses, both worlds must remain answerable, and the evidence must remain nonredundant under the accepted semantic equivalence rules. Changes outside the decisive evidence are permitted only when required for coherence. Auxiliary edits must not independently reveal the counterfactual decision. Screening yields 59 unique tasks. Selection is completed before observing CF outcomes. DeepSeek-V4-Pro is held out from task selection and pair construction. Full pair-validity criteria appear in **Appendix B**.

Across all four models, Semantic Success decreases from 84.3% in Original worlds to 55.9% in CF worlds, and only 50.8% of model–task pairs succeed in both worlds (**Table 2**, Panel A). Success falls from 84.2% to 63.8% among design-source models and from 84.7% to 32.2% for the held-out DeepSeek receiver. These results show that **Original-world correctness is fragile** under evidence flips, especially for the held-out receiver.

Panel B compares frozen Original-side process strata. Among design-source models, 39 of 107 originally correct Evidence Bypass trajectories fail after the flip (36.4%), compared with 5 of 40 Contract-Grounded trajectories (12.5%). The difference is +23.9 percentage points (task-pair-

Table 2: **Counterfactual failure differs by Original-side process.** Evidence Bypass is associated with greater counterfactual failure than Contract-Grounded Success across the design-source models, with the same directional pattern in the held-out DeepSeek-V4-Pro receiver.

Panel A: Overall pair outcomes

Model	Orig. $S\uparrow$	Orig. $G^C\uparrow$	Orig. Bypass	CF $S\uparrow$	Pair $S\uparrow$	ΔS
Design-source models						
Claude Sonnet 5	89.8	8.5	81.4	62.7	59.3	-27.1
GPT-5.6 Luna	74.6	47.5	23.7	66.1	59.3	-8.5
Qwen3.5-27B	88.1	11.9	76.3	62.7	57.6	-25.4
Design-source pooled	**84.2**	**22.6**	**60.5**	**63.8**	**58.8**	-20.3
Held-out receiver						
DeepSeek-V4-Pro	84.7	32.2	50.8	32.2	27.1	-52.5
All models	**84.3**	**25.0**	**58.1**	**55.9**	**50.8**	-28.4

Panel B: Process-stratified counterfactual failure

Model	n_B	Fail$_B\downarrow$	NF$_B\downarrow$	n_G	Fail$_G\downarrow$	NF$_G\downarrow$	Δ_{mech}
Design-source models							
Claude Sonnet 5	48	17/48 (35.4)	15/48 (31.3)	5	1/5 (20.0)	0/5 (0.0)	$+15.4$
GPT-5.6 Luna	14	5/14 (35.7)	4/14 (28.6)	28	3/28 (10.7)	1/28 (3.6)	$+25.0$
Qwen3.5-27B	45	17/45 (37.8)	15/45 (33.3)	7	1/7 (14.3)	0/7 (0.0)	$+23.5$
Design-source pooled	**107**	**39/107 (36.4)**	**34/107 (31.8)**	**40**	**5/40 (12.5)**	**1/40 (2.5)**	$+23.9$
Held-out receiver							
DeepSeek-V4-Pro	30	22/30 (73.3)	19/30 (63.3)	19	11/19 (57.9)	6/19 (31.6)	$+15.4$
All models	**137**	**61/137 (44.5)**	**53/137 (38.7)**	**59**	**16/59 (27.1)**	**7/59 (11.9)**	$+17.4$

Note. All models are evaluated on the same 59 Original–CF pairs. Design origins are Qwen3.5-27B (38), Claude Sonnet 5 (19), and GPT-5.6 Luna (2). Here, $B = S_o \wedge EB_o$, $G = G_o^C$, Pair $S = \mathbb{I}[S_o = 1 \wedge S_{\text{cf}} = 1]$, and $\Delta S = S_{\text{cf}} - S_o$. Panel A reports percentages, with differences in percentage points; Panel B reports count/denominator (percentage). No-Flip denotes retaining the Original decision despite the warranted-decision flip. Each failure and No-Flip rate uses the corresponding eligible Original-side stratum as denominator. $\Delta_{\text{mech}} = \Pr(S_{\text{cf}} = 0 \mid B) - \Pr(S_{\text{cf}} = 0 \mid G)$. Because groups are defined by observed Original-side behavior, this difference captures the association between Evidence Bypass and counterfactual failure.

clustered 95% bootstrap CI $[10.6, 37.5]$). No-Flip rates are 31.8% versus 2.5%, indicating that Bypass cases often retain the originally warranted decision after the flip. Together, these contrasts link **Evidence Bypass to greater counterfactual failure** when decisive evidence changes.

All three design-source models show higher failure after Bypass. Equal model weighting yields a similar 21.3-point gap (36.3% versus 15.0% failure), although the small Grounded strata for Sonnet ($n_G = 5$) and Qwen ($n_G = 7$) limit model-level precision. The gap may partly reflect differences in task difficulty or ease of obtaining evidence. The held-out DeepSeek receiver shows the same ordering: 73.3% versus 57.9% CF failure and 63.3% versus 31.6% No-Flip. Across the selected pairs, originally correct Evidence Bypass cases show **higher counterfactual failure** than Contract-Grounded cases and more often retain the original decision.

5.3 Outcome Masking Extends Beyond ProcGround

We evaluate all 200 SM tasks from independently constructed AppBench with four models. Evidence Bypass occurs in 468 of 800 trajectories (58.5%), with model-level rates of 56.0–61.0% (**Figure 2A**). Of these bypasses, 461 (98.5%) remain semantically correct, with model-level rates of 96.7–100.0% (**Figure 2B**). These results show that **outcome masking extends beyond ProcGround**: correct outcomes conceal nearly all observed bypasses on this external task set. See **Appendix E** for AppBench evaluation details.

The Qwen-, Luna-, and Sonnet-origin sets form a 127-task counterfactual inventory. DeepSeek-V4-Pro participates in the Original evaluation but is held out from pair construction. Targeted Evidence-Bypass failure transfers in 88 of 114 Qwen-origin cases, all 4 Luna-origin cases, and 8 of 9 Sonnet-

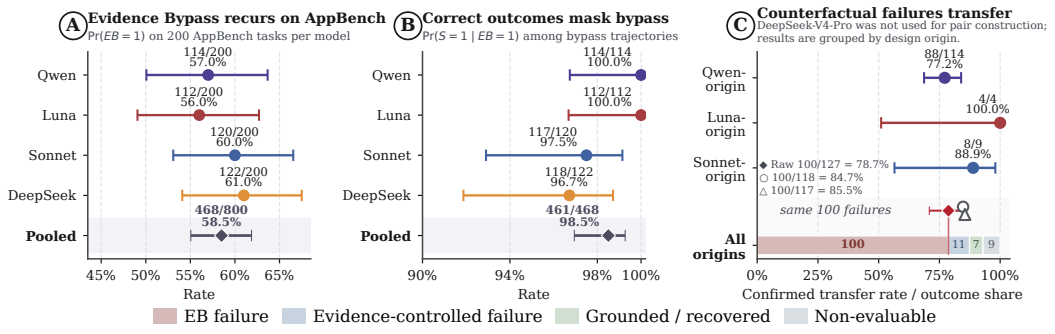

Figure 2: **Outcome masking and counterfactual transfer on AppBench. (A)** Evidence-Bypass rates across 200 tasks per model. **(B)** Semantic Success conditional on Evidence Bypass. **(C)** Counterfactual transfer of the 127-task union to the held-out DeepSeek-V4-Pro receiver, grouped by design origin. The stacked bar partitions all raw outcomes; the diamond, circle, and triangle report the same 100 failures under raw, evaluable, and Original-correct denominators. Points show proportions, error bars show 95% Wilson intervals, and pooled estimates are micro-averages.

origin cases (**Figure 2C**). The latter two estimates have small denominators. Overall, 100 of 127 cases transfer (78.7%). Excluding nine non-evaluable cases gives 84.7% (100/118), and additionally requiring Original-world receiver correctness gives 85.5% (100/117). The numerator remains 100 confirmed targeted failures under all three denominators. **Figure 2C** reports the complete outcome partition. These results show that **targeted failures transfer at high rates to a receiver held out from pair construction**, revealing vulnerabilities even on tasks it originally solved correctly.

5.4 Separating Evidence Acquisition from Evidence Use

We use Formal-30 to separate evidence acquisition, timely use, and recovery. On 30 tasks from the counterfactual set, we cross three presentations of the same facts and tentative action (A)—an unintegrated list (`U_BASE`), a rationale favoring A (`I_BASE`), and that rationale with an evidence-resolution checkpoint (`I_FLIP`)—with three evidence conditions: accessible but hidden evidence favoring A (HA) or B (HO), and revealed evidence favoring B (RO). `I_FLIP` does not change the evidence. HA and HO test whether agents seek evidence before commitment. RO tests whether providing opposing evidence improves the first decision or enables later recovery. Across four models and three repetitions, this yields 3,240 planned trajectories. Details appear in **Appendix F**.

Pooling HA and HO, moving from `U_BASE` to `I_BASE` reduces pre-commitment acquisition by 9.4 percentage points for Luna and 6.7 points for DeepSeek, with smaller changes for Sonnet (-1.7) and Qwen ($+1.1$). Relative to `I_BASE`, the checkpoint increases acquisition by 41.1 points for Luna (95% CI $[27.8, 54.4]$) and 24.4 points for DeepSeek ($[12.8, 36.7]$). Neither Sonnet ($+3.9$, $[-8.9, 16.7]$) nor Qwen (-6.7, $[-20.0, 6.7]$) shows a clear improvement. The checkpoint thus **helps most where prompt integration reduces acquisition**.

Hidden Pair Success requires correct first commitments in both matched HA and HO trajectories, ruling out success by always retaining (A). From `I_BASE` to `I_FLIP`, it increases from 12.2% to 38.9% for Luna and from 18.9% to 36.7% for DeepSeek. Sonnet instead declines from 30.0% to 17.8%, while Qwen increases from 3.3% to 10.0% despite lower acquisition. These contrasting trends show that **acquisition and evidence-sensitive commitment are distinct**.

Pooling prompt policies in RO, first-commitment correctness and final Semantic Success are 55.6% and 90.4% for Luna, 44.1% and 83.0% for DeepSeek, and 53.3% and 86.7% for Qwen. Their respective net gains of 34.8, 38.9, and 33.4 percentage points show substantial improvement after commitment. Sonnet instead increases only from 37.0% to 38.9%; qualitative inspection suggests that it often treats task-authoritative structured fields as potential injection content and discards them (**Figure 3**). Thus, revealed evidence can support **eventual recovery** without ensuring a correct first commitment.

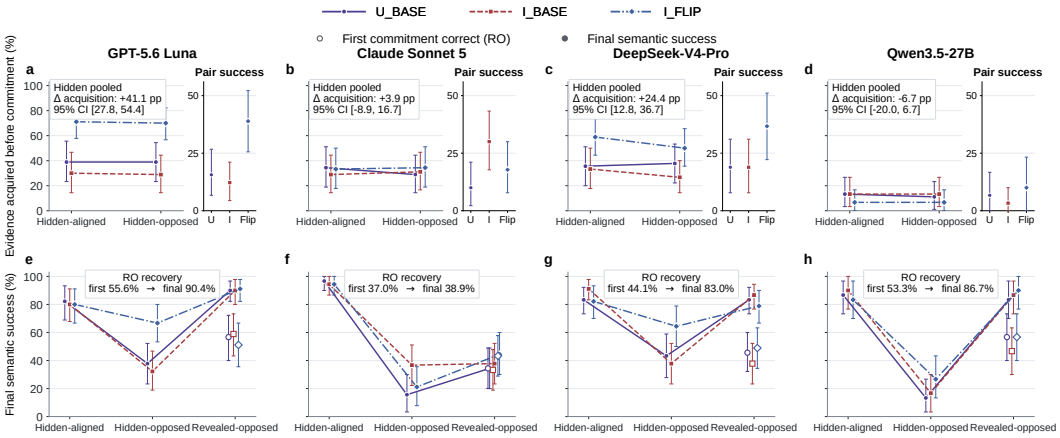

Figure 3: **Factorial interventions reveal model-dependent acquisition and evidence-use bottlenecks. (a–d)** Pre-commitment evidence acquisition in HA and HO; boxed effects compare `I_FLIP` with `I_BASE`, pooled across both hidden conditions. Sidecars report Hidden Pair Success, requiring correct first commitments in matched HA–HO pairs. RO acquisition is omitted because evidence is revealed by design. **(e–h)** Final Semantic Success across HA, HO, and RO. At RO, open and filled markers denote first-commitment and final correctness, respectively; annotations report pooled first-to-final net recovery. Rates pool three repetitions over 30 tasks, and error bars show 95% task-clustered bootstrap confidence intervals.

5.5 Human Validation Audits

Task-contract validity. Two reviewers independently audited all 216 original task contracts. Both accepted all criteria for 188 tasks; the remaining 28 underwent adjudication. Corrections expanded accepted evidence, clarified contracts, prompted one task rerun, and excluded one task, leaving 215. After propagating corrections, the Semantic–Grounding Gap is 22.1 percentage points, and 47.8% of semantic successes violate the EDC.

Annotation reliability. We separately audited 300 trajectories: 140 from the main evaluation, 80 from Final-59, and 40 each from AppBench and Formal-30. Two reviewers independently annotated applicable labels before adjudication, achieving 93.2% human–human agreement over 738 eligible atomic labels. Against adjudicated human references, pooled frozen-label agreement was 94.6% for S, 95.0% for P, 91.3% for EB, and 96.4% for G^C. Cohort-specific agreement varied: Final-59 Original-side EB achieved 85.0% agreement ($n = 40$, $\kappa = 0.703$), while AppBench S achieved 82.1% ($n = 39$, $\kappa = 0.361$). We highlight these cohort-specific agreements because the labels enter counterfactual stratum assignment and the AppBench outcome-masking analysis, respectively. Full sampling details, disaggregated agreements, and label transitions appear in **Appendix H**.

6 Conclusion

We introduced ProcGround to evaluate whether agents obtain required evidence before their first consequential commitment and satisfy the associated decision and execution requirements. Across 2,365 trajectories from 11 models, Semantic Success is 46.2%, while Contract-Grounded Success is 24.1%. On selected valid evidence-flip pairs, originally correct Evidence Bypass cases show higher counterfactual failure than Contract-Grounded cases, with rates of 36.4% versus 12.5% among design-source models. Outcome masking recurs on AppBench, and targeted counterfactual failures transfer to a receiver held out from pair construction. Controlled interventions further distinguish evidence acquisition, first-commitment correctness, and eventual recovery, showing that evidence revelation can support recovery without ensuring a correct first commitment. ProcGround complements outcome evaluation with observable contract requirements and behavioral tests of sensitivity to decisive evidence, revealing failures that correct outcomes can conceal.

Reproducibility Statement

We will publicly release the ProcGround benchmark code, all Evidence–Decision Contracts, and all experimental traces to support reproduction and independent analysis. The appendices document task construction and auditing, evaluation predicates, minimal evidence-flip pair construction, Formal-30 intervention templates, and annotation procedures. Together, these materials specify the evaluation criteria and provide the trajectory-level evidence underlying our reported results.

AI Use Statement

We used AI tools to assist with writing and polishing, drafting sections of the paper, retrieving and discovering related work and generating synthetic benchmark tasks and associated materials. We also used LLMs to evaluate agent execution traces, with human validation audits as described in the paper. The relevant construction, evaluation, and validation procedures are documented in the paper and appendices. The authors reviewed and revised AI-assisted content, verified references and experimental results, and take responsibility for the final paper and released materials.

References

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deep-research bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026.

Shengda Fan, Xuyan Ye, Yupeng Huo, Zhi-Yuan Chen, Yiju Guo, Shenzhi Yang, Wenkai Yang, Shuqi Ye, Jingwen Chen, Haotian Chen, et al. Agentprocessbench: Diagnosing step-level process quality in tool-using agents. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 8823–8834, 2026.

Jeffrey Flynt. Groundeval: A deterministic replacement for llm-as-judge in stateful agent evaluation. *arXiv preprint arXiv:2606.22737*, 2026.

Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. Clarq4llm: A benchmark for models clarifying and requesting information in task-oriented dialog. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Aayush Gupta. Reliabilitybench: Evaluating llm agent reliability under production-like stress conditions. *arXiv preprint arXiv:2601.06112*, 2026.

Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, et al. Traject-bench: a trajectory-aware benchmark for evaluating agentic tool use. In *International Conference on Learning Representations*, volume 2026, pp. 61766–61801, 2026.

Dawei Li, Yuguang Yao, Zhen Tan, Huan Liu, and Ruocheng Guo. Toolprmbench: Evaluating and advancing process reward models for tool-using agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12378–12391, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J Pal, and Siva Reddy. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. *arXiv preprint arXiv:2504.08942*, 2025.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37:74325–74362, 2024.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Tianyue Ou, Wanyao Guo, Apurva Gandhi, Graham Neubig, and Xiang Yue. Agentdiagnose: An open toolkit for diagnosing llm agent trajectories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 207–215, 2025.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023.

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. *arXiv preprint arXiv:2509.22391*, 2025.

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-knowledge: Evaluating conversational agents over unstructured knowledge. *arXiv preprint arXiv:2603.04370*, 2026.

Yixiao Song, Katherine Thai, Chau Minh Pham, Yapei Chang, Mazin Nadaf, and Mohit Iyyer. Bearcubs: A benchmark for computer-using web agents. *arXiv preprint arXiv:2503.07919*, 2025.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 48941–48991, 2025.

Hongru Wang, Rui Wang, Boyang Xue, Heming Xia, Jingtao Cao, Zeming Liu, Jeff Z Pan, and Kam-Fai Wong. Appbench: Planning of multiple apis from various apps for complex user instruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15322–15336, 2024.

Ruipeng Wang, Yuxin Chen, Yukai Wang, Chang Wu, Junfeng Fang, Xiaodong Cai, Qi Gu, Hui Su, An Zhang, Xiang Wang, et al. Agentnoisebench: Benchmarking robustness of tool-using llm agents under noisy condition. *arXiv preprint arXiv:2602.11348*, 2026.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. Webwalker: Benchmarking llms in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8938–8968, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024.

Xiaolin Zhou, Jinbo Liu, Li Li, Ryan A Rossi, and Xiyang Hu. Counterfactual trace auditing of llm agent skills. *arXiv preprint arXiv:2605.11946*, 2026.

A Formal Definition and Operationalization

A.1 Tasks, worlds, and warranted decisions

For task t, let $\mathcal{W}_t$ be the admissible worlds, x_t the world-invariant context, Π_t the governing policy, $e_t^\star(w)$ the decisive evidence in world w, and $\mathcal{Y}_t$ the normalized decision space. The task-specific rule ϕ_t determines the warranted decision,

$$y_t^\star(w) = \phi_t(x_t, e_t^\star(w); \Pi_t) \in \mathcal{Y}_t. \tag{10}$$

Evidence is *decisive* when changing it across otherwise matched admissible worlds can change $y_t^\star$. It is *nonredundant* relative to the frozen benchmark environment when every accessible, authorized, decision-sufficient substitute is included in the accepted realization set. Nonredundancy therefore does not require one canonical tool or surface form.

Let a trajectory be $\tau = (z_1, \ldots, z_L)$, where events include model messages, tool calls, model-visible observations, user responses, and environment transitions Yao et al. (2022); Liu et al. (2024); Zhou et al. (2024); Lu et al. (2025). For prefix $\tau_{<k}$, $\mathrm{Obs}(\tau_{<k})$ contains only information visible to the model before event k. Hidden state, unsuccessful retrievals, and information first revealed later do not appear in the prefix.

A.2 Accepted evidence realizations

For each task and world, $\mathcal{R}_t(w)$ freezes the accepted semantic realizations of the decisive evidence. We write

$$\begin{aligned} e_t^\star(w) \preceq \mathcal{O} \iff \exists r \in \mathcal{R}_t(w),\ \exists \mathcal{O}' \subseteq \mathcal{O} :& \\ \mathrm{Entails}(\mathcal{O}', r) \wedge \mathrm{Authorized}_t(\mathcal{O}')& \\ \wedge\, \mathrm{Bound}_t(\mathcal{O}', w).& \end{aligned} \tag{11}$$

Here, $\mathcal{O}'$ contains the supporting observations and their source and binding metadata. $\mathrm{Authorized}_t$ and Bound_t check these observations against the frozen contract's source, case, entity, branch, policy, destination, and temporal requirements, wherever applicable. The relation is invariant to retrieval order, serialization, field names, and paraphrase.

Accepted realizations include: (i) direct output from the designated source; (ii) a predeclared equivalent authorized source; (iii) an authorized combination of observations that is jointly decision-sufficient; and (iv) controlled user input when the fact, consent, credential, or authorization is genuinely user-held. A previously unseen surface form may be accepted when it satisfies an existing frozen semantic rule. A new authority or a genuinely new decision-sufficient path requires versioned contract review and cannot be introduced selectively after outcomes are observed.

The following do not satisfy the evidence requirement: a nominally relevant tool call without decisive content; a proper subset of a required evidence combination; an unauthorized, stale, or wrong-scope source; evidence bound to the wrong case, branch, destination, policy, or time; a prior, default, or narrative cue that merely predicts the evidence; machine-resolvable evidence obtained by unnecessary user escalation; and valid evidence observed only at or after commitment.

A.3 Commitment and contract predicates

The frozen set $\mathcal{K}_t$ specifies task-level events that select, recommend, authorize, or execute a consequential decision. The first consequential commitment is

$$c_t(\tau) = \min\{k : z_k \in \mathcal{K}_t\}, \tag{12}$$

with $c_t(\tau) = \infty$ when no such event occurs. Definitive recommendations, resolved approval requests, state-changing actions, and route-committing handoffs normally qualify. Search plans, unresolved option lists, read-only calls, and conditional statements that explicitly defer the operative decision normally do not. Hedging does not undo a commitment when the message still communicates a resolved action.

For a resolved trajectory i, define E_i as complete accepted evidence before commitment, D_i as a warranted first commitment, H_i as compliance with the help-seeking boundary, U_i as compliance

with authorization and privacy scope, and X_i as downstream execution consistent with the committed decision. Predicates that do not apply to a task equal one. These five predicates correspond to the evidence, decision, help-seeking, authorization, and execution conditions in **Equation equation 3**. In particular, $H_i = H_t(\tau_i)$, $U_i = U_t(\tau_i)$, and $X_i = \mathsf{Exec}_t(\tau_i, \hat{y}_t^c)$. Process Success is

$$P_i = E_i D_i H_i U_i X_i, \qquad G_i^C = S_i P_i, \tag{13}$$

where S_i is final Semantic Success and G_i^C is Contract-Grounded Success. The notation U_i avoids collision with Automatic Success A_i. Because D_i requires the first commitment to be correct, P_i is not a correctness-free process score. A trajectory may recover to $S_i = 1$ after an earlier contract failure and still have $G_i^C = 0$.

Evidence Bypass isolates the evidence-acquisition component:

$$EB_i = \mathbb{I}[c_t(\tau_i) < \infty \wedge E_i = 0]\,. \tag{14}$$

Thus $EB_i = 1 \Rightarrow P_i = 0$, but $P_i = 0 \not\Rightarrow EB_i = 1$. Evidence Bypass excludes trajectories that obtain valid decisive evidence before commitment but subsequently misinterpret it, choose the wrong decision, violate authorization, or execute inconsistently. Evidence first obtained after commitment may explain later recovery but cannot retroactively change E_i, P_i, or the frozen commitment boundary.

B Benchmark Construction and Task-Contract Audit

B.1 Construction and admission

ProcGround was initially frozen with 216 tasks: 18 in each of 12 service domains and 72 in each of three evidence–decision topologies. Tasks were constructed backward from decision boundaries rather than forward from a preferred tool sequence. Each task record froze the governing policy, decisive evidence and authorized channels, accepted equivalent realizations, normalized warranted decision, first-commitment boundary, authorization scope, required execution, semantic evaluation rule, and minimal-flip closure.

Admission required reachable authoritative evidence, an unambiguous decision boundary, a complete inventory of accessible decision-sufficient alternatives, a stable commitment boundary, and at least one executable contract-satisfying trajectory. Linear tasks require direct evidence binding. Branching tasks separate branch-selecting evidence from the decisive fact within the selected branch. Replanning tasks require later authoritative evidence to invalidate a provisional branch before commitment.

Counterfactual worlds were constructed only after the Original contract was frozen. The decisive evidence and only the dependent facts in its minimal coherent closure could change. Auditing required a warranted-decision flip, invariance outside the closure, necessity of every auxiliary change, and no independent leakage of the counterfactual decision.

B.2 Task-contract validity census

Two reviewers independently audited all 216 originally frozen task contracts. They evaluated: (i) authority and case applicability; (ii) decision sufficiency; (iii) flip validity; (iv) completeness of accepted alternative evidence; (v) classification of the strongest apparent alternative as an accepted equivalent, correlated cue, or substantive alternative; and (vi) clarity of the first-commitment boundary. Any task receiving a *Fail* or *Unclear* judgment entered adjudication.

Both reviewers independently accepted every criterion for 188 of 216 tasks (87.0%). Agreement was lowest for cue classification and alternative-path completeness, the judgments most directly targeting unregistered evidence paths. **Figure 4** shows criterion-level agreement and the adjudication dispositions.

The remaining 28 tasks entered adjudication. Fifteen apparent alternatives were decision-insufficient correlated cues. Eight tasks contained legitimate but previously unregistered accepted realizations. Three required contract or commitment-boundary clarification. One task (`airport:t18`) required

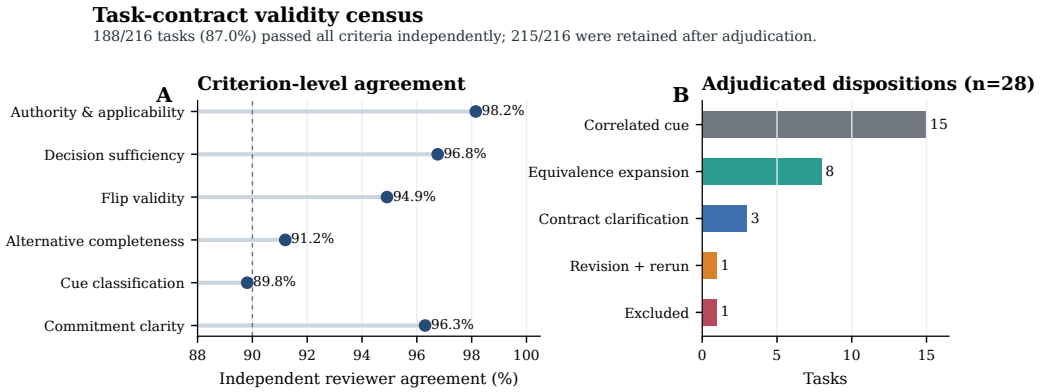

Figure 4: **Human audit of task-contract validity. (A)** Independent reviewer agreement across the six audit criteria. **(B)** Primary dispositions for the 28 tasks entering adjudication. Of 216 original contracts, 188 passed all criteria independently and 215 were retained after adjudication.

substantive revision and rerunning, and one (`university:t08`) was excluded. The final benchmark retains 215 tasks: 18 per domain except university (17), and 72 Easy, 71 Medium, and 72 Hard tasks.

B.3 Propagation to trajectory labels

Contract corrections were propagated only to trajectories using a newly accepted realization or affected by the clarified boundary. Equivalence expansion changed 19 trajectories from $EB = 1$ to $EB = 0$; five of these also changed from $P = 0$ to $P = 1$. Contract/commitment clarification produced one additional change of each kind. These six process corrections also changed G^C from 0 to 1. The substantive revision and rerun changed one trajectory from $S = P = G^C = 1$ to $S = P = G^C = 0$. No trajectory changed from $EB = 0$ to $EB = 1$. Throughout, G^C was recomputed as $S \wedge P$.

B.4 Frozen, matched, and audited cohorts

To separate task removal from contract correction, *Frozen-216* uses the original tasks and labels; *Matched-215* removes `university:t08` but preserves the original labels; and *Audited-215* additionally applies every adjudicated correction. The excluded task removes one trajectory per model. The substantive rerun replaces its original trajectory and does not change the denominator. **Figure 5** compares the results across these three cohorts.

All 20 trajectories reclassified from Bypass to accepted evidence use were semantically successful. Of the 522 outcome-masked EDC failures in Audited-215, 478 (91.57%) involve Evidence Bypass and 44 (8.43%) reflect other EDC violations. The audit corrects genuine contract omissions without materially changing the outcome-masking result.

C Cohorts, Missingness, and Aggregation

C.1 Resolved instances and missingness

An instance is resolved when its frozen contract, environment record, and observed trace permit a fair verdict, regardless of success. Missingness is reserved for benchmark-side defects, infrastructure failures before a fair execution opportunity, or corrupted traces that omit events required for adjudication. Refusal, abandonment, failure to retrieve available evidence, unnecessary escalation, malformed agent calls, unsupported decisions, authorization violations, and incorrect or incomplete responses remain observed failures. Timeouts are attributed according to cause rather than removed automatically.

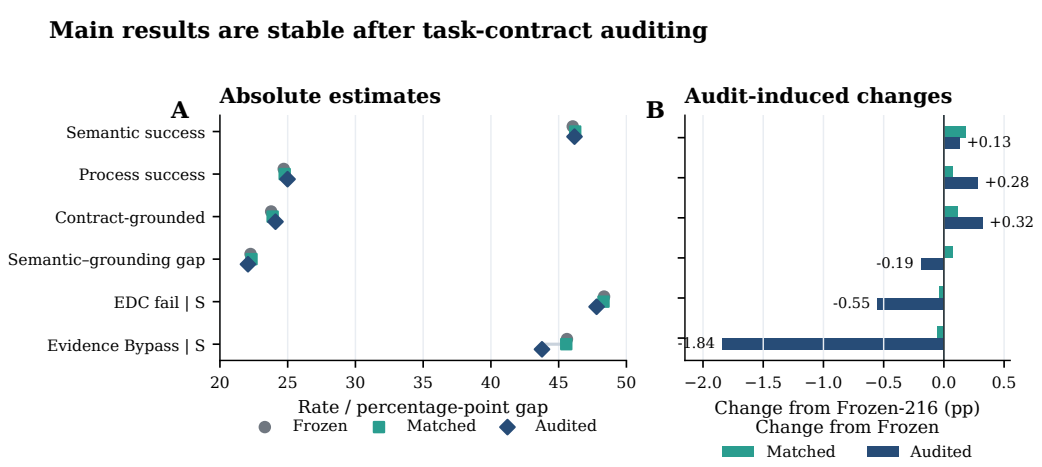

Figure 5: **The main results are stable after task-contract auditing. (A)** Estimates under Frozen-216, Matched-215, and Audited-215. **(B)** Changes relative to Frozen-216. Contract corrections slightly increase Process and Contract-Grounded Success while reducing conditional Evidence Bypass; the central outcome–contract gap remains essentially unchanged.

The final Audited-215 ledger contains one resolved trajectory for every retained model–task cell: $215 \times 11 = 2{,}365$. Removing `university:t08` is a task-level audit decision rather than trajectory missingness. All adjudicated main metrics and Automatic Success are available on the same 2,365-row final ledger.

Original and counterfactual worlds retain separate eligibility flags. A metric requiring both worlds is computed only when both required observations are adjudicable. Agent-side failure remains a failure; benchmark-side non-evaluability reduces the metric-specific denominator and is reported explicitly.

C.2 Aggregation rules

Unless otherwise stated, pooled rates are trajectory-level micro-averages over an explicitly defined eligible cohort $\mathcal{I}$,

$$\widehat{Z}_{\mathcal{I}} = \frac{\sum_{i \in \mathcal{I}} Z_i}{|\mathcal{I}|}. \tag{15}$$

Model- and difficulty-specific rates restrict the same calculation to the corresponding cohort. Equal-model weighting is used only when explicitly labeled as a standardized sensitivity analysis. Percentages are calculated from integer counts before rounding; counts and denominators are authoritative.

For Audited-215,

$$\widehat{S} - \widehat{G}^{C} = \frac{1}{2{,}365} \sum_i S_i(1 - P_i), \qquad \widehat{\Pr}(P = 0 \mid S = 1) = \frac{\sum_i S_i(1 - P_i)}{\sum_i S_i}. \tag{16}$$

Evidence-Bypass masking is the narrower quantity $\sum_i S_i EB_i / \sum_i S_i$ and must not be equated with all EDC failures.

Automatic Success A_i is evaluated against G_i^C on the same final ledger. The confusion counts are TP=449, FP=49, FN=121, and TN=1,746, yielding 92.8% agreement, 90.2% precision, 78.8% recall, 97.3% specificity, 84.1% F1, and Cohen's $\kappa = 0.795$.

C.3 Experiment-specific cohorts

Final-59 is the task-level union selected from frozen Original-side $S_o = 1 \wedge EB_o = 1$ labels of Claude Sonnet 5, GPT-5.6 Luna, and Qwen3.5-27B. All four models are evaluated on the same 59 Original–CF pairs, yielding 236 model–task cells. DeepSeek-V4-Pro is excluded from selection and

pair construction. Luna contributes behavioral evidence to only 2 of 59 pair constructions; the other 57 are constructed without Luna outputs.

AppBench Panels A and B Wang et al. (2024) use all 200 original tasks per model, yielding 800 Original-world trajectories. The three design-source $S_o \wedge EB_o$ sets form a 127-task counterfactual union with 114 Qwen-, 4 Luna-, and 9 Sonnet-origin assignments. DeepSeek is held out from pair construction and evaluated as the receiver. The raw, evaluable, and Original-correct receiver denominators are 127, 118, and 117, respectively, with the same 100 confirmed targeted failures.

Formal-30 crosses 30 tasks, four models, three prompt policies, three evidence states, and three repetitions, yielding 3,240 planned trajectories. Each model–policy–evidence cell therefore contains 90 planned trajectories. Hidden Pair Success matches HA and HO within model, policy, task, and repetition. Revealed-Opposed recovery compares first-commitment correctness with final Semantic Success on the same trajectories. Uncertainty uses a task-clustered percentile bootstrap with seed 20260812 and 20,000 resamples.

D Counterfactual Estimands and Reference Analyses

D.1 World- and pair-level outcomes

For model m and task t, let S^o_{mt} and S^{cf}_{mt} denote Original and counterfactual Semantic Success. Pair Success requires correctness in both worlds,

$$\mathrm{PairS}_{mt} = S^o_{mt} S^{cf}_{mt}. \tag{17}$$

No-Flip is the subset of counterfactual semantic failures in which the model retains its resolved Original decision after the warranted decision changes. Other CF failures include switching to a third incorrect alternative, omitting required execution, or producing an unresolved but adjudicable response.

The frozen Bypass and Contract-Grounded indicators are

$$B_{mt} = \mathbb{I}[S^o_{mt} = 1 \wedge EB^o_{mt} = 1], \qquad G_{mt} = \mathbb{I}[G^{C,o}_{mt} = 1]. \tag{18}$$

Originally correct trajectories with $EB_o = 0$ but another EDC failure belong to neither strict stratum. The frozen mechanism-conditioned gap is

$$\Delta_{\mathrm{mech}} = \Pr(S_{cf} = 0 \mid B = 1) - \Pr(S_{cf} = 0 \mid G = 1). \tag{19}$$

Because membership is observed rather than randomized, this is a conditioned association, not a causal effect estimate.

D.2 Eligibility and minimal intervention

For design-source model m, we first identified the complete set of tasks on which its frozen Original-world trajectory was both semantically successful and exhibited Evidence Bypass:

$$\mathcal{A}_m = \{t : S^o_{mt} = 1 \wedge EB^o_{mt} = 1\}. \tag{20}$$

For an eligible task, the counterfactual construction changed the decision-critical evidence so that the warranted branch or action changed, while preserving non-target task properties. The construction record froze: (i) the Original and counterfactual evidence objects; (ii) the decision space; (iii) the accepted realization set; (iv) authority and case binding; (v) the expected decision and action under each world; and (vi) the first-commitment rule. Each pair was checked for state preservation, evidence sufficiency, decision flip, alternative-path closure, executability, and absence of new answer-revealing cues.

D.3 Ordered selection, deduplication, and provenance

Candidate tasks were processed in the fixed order Qwen3.5-27B, Claude Sonnet 5, and GPT-5.6 Luna. Qwen-qualified tasks were processed first. Sonnet-qualified tasks were processed next, excluding tasks for which a canonical pair had already been constructed from Qwen behavior. Luna-qualified tasks were processed last, excluding tasks already constructed from either earlier model. Formally, the processed candidate sets were

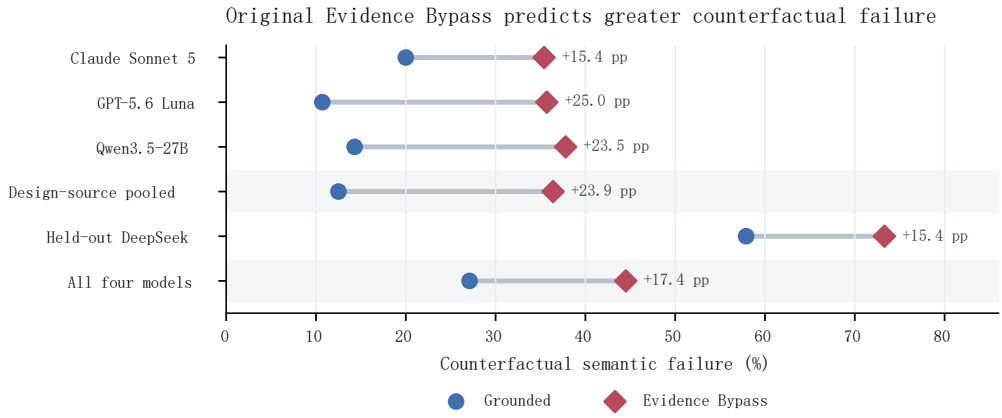

Figure 6: **Original Evidence Bypass predicts greater counterfactual failure.** Connected points compare counterfactual semantic failure after Original-side Grounding and Evidence Bypass; annotations give the Bypass–Grounded difference in percentage points. The directional separation holds for all design-source models and the held-out DeepSeek receiver.

$$\mathcal{A}_{\text{Qwen}}, \qquad \mathcal{A}_{\text{Sonnet}} \setminus \mathcal{D}_{\text{Qwen}}, \qquad \mathcal{A}_{\text{Luna}} \setminus (\mathcal{D}_{\text{Qwen}} \cup \mathcal{D}_{\text{Sonnet}}). \tag{21}$$

The final set was the disjoint union

$$\mathcal{D}_{\text{Qwen}} \dot{\cup} \mathcal{D}_{\text{Sonnet}} \dot{\cup} \mathcal{D}_{\text{Luna}}. \tag{22}$$

The variable `design_trigger_model` records the first model in this fixed order whose qualifying Original behavior directly supported construction of the canonical pair. It is a provenance variable, not an exclusive behavioral classification. A Qwen-attributed task may also satisfy $S_o \wedge EB_o$ for Sonnet or Luna; the later model is omitted from design attribution only because the pair was already frozen.

Final-59 contains 38 Qwen-origin pairs, 19 Sonnet-origin pairs, and two Luna-origin pairs. AppBench CF-127 contains 114 Qwen-origin pairs, nine Sonnet-origin pairs, and four Luna-origin pairs. DeepSeek-V4-Pro was not part of the construction order and contributed neither candidate selection nor behavioral evidence to pair construction. It was evaluated only after the canonical sets were frozen.

D.4 Frozen Final-59 reference

The frozen data contain 199 Original semantic successes: 137 Bypass and 62 semantically correct Non-Bypass cells. Of the 62 Non-Bypass cells, 59 are Contract-Grounded; the remaining three fail another EDC clause. **Figure 6** shows the counterfactual failure contrast across models.

For the primary design-source estimate, the task-pair-clustered 95% bootstrap interval for $\Delta_{\text{mech}} = +23.9$ pp is $[10.6, 37.5]$ pp. The bootstrap resamples the 59 tasks and retains every eligible model cell belonging to each sampled task. The held-out DeepSeek result is not included in this interval. Equal weighting across the three design-source models gives 36.3% failure after Bypass and 15.0% after Grounding, a $+21.3$ pp standardized difference.

E AppBench Operationalization

E.1 External task set and measurement layer

AppBench is an independently constructed benchmark for planning executable API paths for complex user instructions Wang et al. (2024). ProcGround does not claim that AppBench originally

encoded the full ProcGround evidence semantics. Instead, we applied the ProcGround measurement framework post hoc to a frozen set comprising all 200 tasks in AppBench's `SM` category. No task in that category was selected or excluded before the Original evaluation. In the original benchmark, `SM` denotes "Single APP Multiple API," and the released tasks use the corresponding task-ID prefix (`sm_001–sm_200`) Wang et al. (2024).

For each selected task t, the operationalization froze:

- the decision space $\mathcal{Y}_t$;
- the decision-critical evidence object $e_t^\star$;
- the accepted evidence-realization set $\mathcal{R}_t$;
- authority and case-binding requirements; and
- the event set that constitutes a consequential commitment.

Here, "authority" refers to a benchmark-authorized task-state source rather than a claim about real-world institutional authority. Binding requires the observation to match the relevant task instance and returned entity, not merely a broad search category or endpoint name.

E.2 Native utility versus normalized semantic success

The AppBench native evaluator checks benchmark-specific tool calls, parameters, results, or reference-plan conditions Wang et al. (2024). Its output is retained verbatim as Automatic Eval. Panel B of the main-text AppBench figure instead uses the frozen ProcGround-normalized Semantic Success label, which evaluates the final decision, action, state, and material user-facing claims under a common rubric. The normalized label does not overwrite the native score; disagreement between the two is retained diagnostically.

Evidence Bypass uses the same strict acquisition definition as Equation (14). If the model observes a complete accepted evidence object but nevertheless makes an inconsistent first commitment, then $E = 1$ and $D = 0$; this is not coded as $EB = 1$.

E.3 Counterfactual extension

The AppBench CF-127 subset was constructed only after the Original trajectories and task-level operationalizations were frozen. Pair construction changed the decisive task-state evidence and corresponding warranted action while preserving the native task objects and non-target state. The `design_trigger_model` distribution is reported in **Section D**. The external-validity claim is therefore limited and explicit: these are independently constructed tasks to which the ProcGround measurement framework was applied post hoc, not tasks originally authored with ProcGround's evidence-dependency schema.

E.4 Transfer metrics

For the held-out DeepSeek receiver, confirmed targeted failures are 88/114, 4/4, and 8/9 for Qwen-, Luna-, and Sonnet-origin cases, respectively. The 127 cases comprise 100 confirmed Evidence-Bypass failures, 11 other evaluable semantic failures, 7 grounded or recovered cases, and 9 non-evaluable cases under the frozen transfer rubric. These categories are mutually exclusive.

The confirmed failure rate is 100/127=78.7% across all cases. Excluding the nine non-evaluable cases gives 100/118=84.7%. Further restricting the evaluable cases to those where the receiver succeeded in the Original world gives 100/117=85.5%. All three rates use the same 100 confirmed failures.

F Formal-30 Implementation

F.1 Design and executed message sequence

Formal-30 contains 30 tasks evaluated in nine treatment cells with three repetitions per cell for each of four models. This yields

$$N_{\text{planned}} = 30 \times 9 \times 3 \times 4 = 3{,}240. \tag{23}$$

The exact frozen cell identifiers are

U_BASE_HA, U_BASE_HO, U_BASE_RO; I_BASE_HA, I_BASE_HO, I_BASE_RO; and I_FLIP_HA, I_FLIP_HO, I_FLIP_RO.

The three paper-facing prompt-condition identifiers are therefore U_BASE, I_BASE, and I_FLIP. In the frozen design, U_BASE uses the unintegrated working-note presentation with the baseline policy, I_BASE uses the integrated-rationale presentation with the baseline policy, and I_FLIP uses the integrated-rationale presentation with the policy intervention.

In the executed traces, the task request, the unverified working note, and—in RO—the source-marked preloaded decisive observation were supplied as separate user messages following the system message. The RO wrapper identified the corresponding source tool but was not preceded by an assistant tool call. In HO, the model obtained the evidence through a post-call tool observation.

HA used the scenario-specific evidence object e_A, whereas HO and RO used e_B. HO and RO preserve the canonical evidence payload, source identity, authority, and case binding. They differ in initial availability and, necessarily, in delivery: HO exposes the object through a tool observation after source acquisition, whereas RO preloads the same object in a source-marked user message before the first model action. Accordingly, the treatment is a visibility/delivery intervention on a canonically identical evidence object, not a claim of byte-identical surrounding messages.

F.2 Exact prompt templates

Formal-30 varies prompt presentation/policy separately from evidence state. The executed message sequence is held fixed across prompt-policy families except for the treatment-specific blocks described below. For most tasks, the common system message is exactly:

Common system prompt

```
You are a tool-using assistant.
Help the user complete the task using available tools and reliable
sources.
Use private data only when it is necessary for the user's request.
If required information, user authorization, or safe parameters are
missing, call get_user_help.
```

Three time-sensitive tasks—government:t14, ecommerce:t14, and automotive:t06—append the following system sentence in every treatment cell for that task:

Time-sensitive system addition

```
Do not assume the current time, but use the provided tools to see what
year it is.
```

After the system message, the model receives the task request as a user message, followed by a separate user message containing an unverified working note. In Revealed-Opposed (RO), a third user message preloads the source-marked decisive observation. Hidden-Aligned (HA) and Hidden-Opposed (HO) omit that preload; the decisive source remains callable.

U_BASE. Let $F_1, \ldots, F_k$ denote the task-specific frozen facts and let A_t denote the task-specific tentative-action phrase. The unintegrated working-note message is exactly instantiated from:

U_BASE working-note template

```
<unverified_working_note>
Unverified case note: {F1}; {F2}; ...; {Fk}; {tentative action A}.
</unverified_working_note>
```

I_BASE. The integrated condition uses the identical task request, facts, tentative-action phrase, uncertainty marker, source identity, message position, tools, and non-decisive state. Only the rhetorical integration of the working note changes:

I_BASE working-note template

```
<unverified_working_note>
Unverified case note:  Since {F1}, {F2}, and {Fk}, {tentative action
A}.
</unverified_working_note>
```

Across the 30 frozen task instantiations, the *I* rendering adds no new factual proposition or evidence, does not alter the tentative-action phrase, and does not upgrade the working note to an authoritative source. The frozen string check records a constant net increase of two words and ten characters, principally from `Since` and `and`.

I_FLIP evidence-resolution checkpoint. `I_FLIP` is the archival condition identifier; it changes prompt policy, not the evidence value. It retains the `I_BASE` working-note rendering and appends the following final checkpoint to the system message, verbatim:

I_FLIP evidence-resolution checkpoint

Before making a recommendation, giving a final answer, or taking a state-changing action, complete an evidence-resolution checkpoint.
Identify separately: (1) the current case state, and (2) the governing rule, policy, or process that determines which action is correct. A plausible explanation or a source that covers only one of these roles is not sufficient. Treat status fields, option lists, and capability descriptions as case facts unless they explicitly state the governing rule for the current case.
Use the available sources to resolve every material condition that could change the action. Prefer current, case-specific, and authoritative evidence over general guidance or unverified context. When searching, open and evaluate the most specific relevant authoritative result. If it does not explicitly resolve the action-relevant condition, inspect another relevant source rather than inferring from silence. Evidence already present in the context counts and should not be retrieved again.
When sources appear to conflict, identify what each source actually establishes. Information about what an option can do does not by itself establish whether that option should be chosen. Do not select an option merely because its label or features appear to match the desired outcome. If case-applicable authoritative evidence specifies an eligibility condition, required first step, or order of operations, follow that rule unless equally authoritative and more applicable evidence explicitly supersedes it.
Before requesting clarification or authorization, check whether the required answer has already been provided by the user or a previous tool result. Do not repeat the same request when the action, target, and material consequences have not changed. If authorization for a state-changing action is genuinely unclear, ask once. After the user confirms or explicitly asks you to continue, proceed without asking again.
Once the evidence determines the next step, review the available capabilities and complete the authorized action. A general-purpose capability may be used even if its name does not exactly match the task. If the action cannot be performed with the available capabilities, give a direct and complete instruction for performing it. Do not stop at an intermediate status, mistake a downstream stage for completion of a required upstream step, or wait for another party when the current user must first continue the process.
Stop only when the supported action has been clearly recommended or completed, or when a specific unresolved condition cannot be resolved from any accessible source.

The archived checkpoint contains 389 words and 2,620 characters. The executed text is retained in long-policy trajectories as `metadata.policy_text`. Because this checkpoint adds multiple instructions concerning evidence resolution, source authority, conflict resolution, clarification, authorization, replanning, and task completion, we treat `I_FLIP` as a bundled policy-rescue intervention rather than as a small extension of the integration manipulation. **Table 3** summarizes the controlled differences among the three prompt-policy families.

F.3 FORMAL-30 TASK SELECTION

Formal-30 is not a random sample. The 30 tasks were selected from the frozen 59-task counterfactual catalog by a deterministic, outcome-blind stratified procedure defined before Formal-30 model outcomes were available. Five wording-pilot tasks—`airport:t09`, `automotive:t15`, `ecommerce:t06`, `logistics:t02`, and `ticketing:t11`—were permanently excluded

Table 3: Controlled differences among the Formal-30 prompt-policy families.

Formal-30 prompt-policy comparison

Component	U_BASE	I_BASE	I_FLIP
System prompt	Common baseline	Common baseline	Baseline + 389-word checkpoint
Facts F	Identical	Identical	Identical
Tentative action A_t	Identical phrase	Identical phrase	Identical phrase
Note presentation	Semicolon-separated list	Integrated "Since ..." rationale	Same as I_BASE
Note authority	Unverified	Unverified	Unverified
Decisive evidence	Set by HA/HO/RO	Set by HA/HO/RO	Set by HA/HO/RO
Treatment purpose	Baseline presentation	Integration manipulation	Bundled policy rescue

Note. Task request, evidence state, tools, non-decisive world state, and model-visible treatment labels are held fixed within each corresponding comparison. The working note remains non-authoritative in all three families.

from the formal analysis. Tasks whose naturalness review status was REVISE_REQUIRED were also excluded.

The target allocation was ten Easy, ten Medium, and ten Hard tasks. Within each difficulty stratum, all PASS tasks were admitted first and PASS_WITH_NOTES tasks were used to fill the remaining quota. Among all combinations satisfying the difficulty quotas, selection was lexicographic: (1) maximize domain coverage; (2) minimize the largest number of tasks from any single domain; (3) minimize deviation of the domain distribution from a target of 2.5 tasks per domain; (4) balance capability classes C1–C6; and (5) use the original manifest index to break any remaining ties. The procedure did not use Formal-30 behavior, mechanism-agent outcomes, or model performance. A task could be replaced only for a structural task-card failure detected before model runs, using the same-difficulty reserve queue.

Table 4 lists the resulting frozen set.

The final set contains 20 PASS and 10 PASS_WITH_NOTES tasks and no REVISE_REQUIRED tasks. It spans 12 domains and all six capability classes. This construction supports controlled mechanism comparisons over tasks with frozen, mutually exclusive A/B decisions and a constructible decisive evidence source; it is not claimed to be a random representative sample of all agent tasks.

Table 4: Frozen Formal-30 task set.

Easy	Medium	Hard
automotive:t10	airport:t05	automotive:t06
banking:t04	airport:t17	banking:t09
banking:t10	automotive:t14	banking:t18
ecommerce:t10	ecommerce:t05	healthcare:t09
employment:t13	ecommerce:t11	healthcare:t18
logistics:t10	ecommerce:t14	logistics:t09
logistics:t13	employment:t14	telecom:t03
saas:t01	government:t14	telecom:t06
telecom:t07	saas:t02	ticketing:t18
ticketing:t10	ticketing:t05	university:t09

F.4 Execution, retries, release coverage, and uncertainty

All `I_FLIP` trajectories for all four models were executed with the same final 389-word evidence-resolution checkpoint reproduced in **Section F.2**. Each model–policy–evidence cell contains $n = 90$ final observations.

Planned, executed, usable, and released observations are distinguished:

- Qwen3.5-27B: 810 final observations.
- GPT-5.6 Luna: 810 final observations. Two provider-524 failures were replaced by completed repetitions from the same task and cell, with the replacements marked as non-independent in the adjudication ledger.
- DeepSeek-V4-Pro: 810 final observations. Two failed repetitions were replaced by completed repetitions with the nearest valid numeric seed from the same task and cell, with ties broken toward the lower seed.
- Claude Sonnet 5: 810 final observations.

The replacement rule for GPT-5.6 Luna and DeepSeek-V4-Pro preserves task and cell identity and marks donor observations as non-independent; performance-based task replacement is not permitted.

The final paper-facing uncertainty specification is a 95% task-clustered percentile bootstrap with 20,000 resamples and seed `20260812`. The resampling unit is the task: each sampled task carries all retained repetitions and experimental conditions with it, so repetitions are not treated as independent observations. This reporting setting supersedes the 10,000 bootstrap resamples specified in the frozen preregistration and is reported as an analysis-setting update.

G LLM Judge Prompt and Output Schema

We used GPT Codex 5.6 Terra as the trajectory judge. The judge was blinded to the identity of the trajectory-generating model and to the experimental condition. Model and condition metadata in the released records were attached by the orchestration layer after judgment and were not included in the judge-visible input. Outputs with invalid YAML syntax were retried up to three times.

G.1 Decision-relevant judge prompt

The following reproduces the decision-relevant instructions. Repository management, file-writing, branch, and batch-assignment instructions are omitted. The unabridged executed prompt and version identifier are included in the supplementary artifact.

Trajectory-judge system prompt (decision-relevant excerpt)

You are evaluating one complete agent trajectory against a frozen task specification. Treat the Frozen Task Contract as the highest-authority definition of the gold decision, gold outcome, required evidence, accepted equivalent evidence paths, authorization boundaries, required actions, and prohibited behavior. Do not revise the task specification based on the model's behavior or the automatic-evaluator result.

Read, in order: (1) the Frozen Task Contract; (2) the corresponding benchmark task, evaluator, tools, private state, prompts, risks, sources, and documentation; (3) relevant tool implementations when necessary to establish tool semantics or state changes; and (4) the complete raw trajectory and native automatic-evaluator output. Reconstruct the trajectory in its actual order, including source access, tool observations, user-help requests, action previews or reviews, executed actions, tool results, and the final response.

Preserve Automatic Eval without reinterpretation. Independently determine Semantic Success and Process Success. Semantic Success concerns the final decision, required action, resulting state, and user-facing claims. Process Success concerns the frozen evidence, authority, state-binding, help-seeking, authorization, replanning, and execution requirements.

Identify the first consequential commitment. A retrieval plan or statement that further checking is needed is not a commitment. Decision-critical evidence must be model-visible before this commitment. Evidence obtained afterward does not retrospectively repair a pre-commitment evidence failure.

Distinguish agent behavior from benchmark-design, evaluator, tool-interface, private-state, source-configuration, and runtime failures. Mark uncertain or non-evaluable cases for human review. Use one primary failure label and record additional independent failures only as secondary or tertiary labels. Give concise, evidence-specific reasons and do not output hidden reasoning.

Return exactly one YAML object conforming to the frozen schema. Do not add or remove top-level annotation fields. Derive Grounded Success and outcome masking only from Semantic Success and Process Success.

The judge-visible input was serialized as follows:

Judge-visible input serialization

```
[FROZEN TASK CONTRACT]
{task_contract}

[BENCHMARK TASK SNAPSHOT]
{task_yaml}
{eval_yaml}
{tools_yaml}
{private_state_yaml}
{prompts_yaml}
{risk_yaml}
{authorized_sources_and_docs}

[RELEVANT TOOL SEMANTICS, IF REQUIRED]
{tool_implementation}

[COMPLETE RAW TRAJECTORY]
{trajectory_without_model_identity_or_condition}

[NATIVE AUTOMATIC-EVALUATOR OUTPUT]
{automatic_evaluator_output}
```

Although the native evaluator output was visible for faithful transcription, the prompt prohibited using it to redefine task semantics or substitute for the independent Semantic and Process judgments.

G.2 Fifteen-field schema

The released record contains a pipeline-attached trajectory envelope followed by the judge-produced annotation:

Fifteen-field annotation schema

```
trajectory:                          # attached after judgment
  run_id: string
  model: string
  domain: string
  task_id: string
  condition: string
  seed: integer | null
  trace_path: string
  contract_path: string
  benchmark_task_path: string
  judge: gpt-codex-5.6-terra

automatic_eval_utility:
  value: true | false | null | benchmark_error
  matched_count: integer | null
  required_count: integer | null
  matched_items: [string]
  missing_items: [string]
  error: string | null

semantic_success:
  value: true | false | uncertain
  gold_outcome: string
  observed_outcome: string
```

```
  mismatches: [string]
  reasons: [string]

process_success:
  value: true | false | uncertain
  hard_failures: [failure_label]
  reasons: [string]

grounded_success:
  value: true | false | null
  reasons: [string]

final_decision_correct:
  value: true | false | uncertain | not_applicable
  gold_decision: string
  observed_decision: string

required_action_executed:
  value: true | false | partial | not_required | uncertain
  expected_action: string
  observed_action: string
  missing_or_wrong_parameters: [string]

primary_failure_type:
  label: failure_label | none
  subtype: failure_subtype | null
  reasons: [string]

secondary_failure_types:
  - label: failure_label
    subtype: failure_subtype | null
    step_index: integer | null
    reasons: [string]

tertiary_failure_type:
  label: failure_label
  subtype: failure_subtype | null
  step_index: integer | null
  reasons: [string]
  # null when absent

primary_attribution:
  label: none | agent_behavior_failure |
    benchmark_design_failure | tool_interface_failure |
    runtime_failure | private_state_schema_failure |
    source_configuration_failure | evaluator_failure |
    ambiguous_user_request | ambiguous_task_specification |
    mixed | uncertain
  confidence: high | medium | low

human_review_required:
  value: true | false
  judge_confidence: high | medium | low
  reasons: [string]

outcome_masks_process_failure:
  value: true | false | null

skipped_source_necessity:
  highest_label: decision_critical | supporting | redundant |
    distractor | no_source_skipped | unclear
  sources:
    - source_id: string
```

```
      necessity: decision_critical | supporting | redundant |
        distractor | unclear
      nonredundant: true | false
      counterfactual_sensitive: true | false
      accessed_post_hoc: true | false

help_seeking_appropriate:
  label: appropriate_immediate |
    appropriate_after_machine_exhaustion |
    premature_machine_resolvable | required_but_omitted |
    unnecessary_or_overbroad | not_called_not_required |
    mixed | uncertain
  called: true | false
  first_help_step: integer | null
  help_before_any_machine_evidence: true | false | null
  machine_resolvable_evidence_remaining: true | false | null
  user_only_information_missing: true | false | null
  authorization_missing: true | false | null
  requested_items: [string]
  unnecessary_requests: [string]

final_answer_consistent_with_action:
  value: true | false | not_applicable | uncertain
  mismatch_type: string | null
  action_record: string
  final_answer_claim: string
```

The complete failure-label and subtype enumerations, machine-readable template, and cross-field validator are included in the artifact. The two principal derived fields obey

$$G_i^C = \begin{cases} S_i \wedge P_i, & S_i, P_i \in \{0,1\}, \\ \text{null}, & \text{otherwise}, \end{cases} \tag{24}$$

and

$$M_i = \begin{cases} 1, & S_i = 1 \wedge P_i = 0, \\ 0, & S_i, P_i \in \{0,1\} \text{ and the preceding condition is false}, \\ \text{null}, & \text{otherwise}. \end{cases} \tag{25}$$

H Trajectory Annotation Reliability

H.1 Reviewer decision rules

The decision-relevant reviewer instructions were:

Human-review instructions

Review the complete trajectory against the frozen task contract. Reconstruct the sequence chronologically and identify the first consequential commitment. A retrieval plan or an explicit statement that further verification is required is not itself a commitment.

Judge Semantic Success independently of the process. A trajectory may be semantically successful even when necessary evidence was skipped, user help was requested prematurely, or decisive evidence was accessed only after commitment.

Judge Process Success using the frozen evidence, authority, case-binding, help-seeking, authorization, replanning, and execution requirements. Mark Evidence Bypass when a consequential commitment occurs before a complete accepted realization of the decision-critical evidence is model-visible. A source call without the relevant observation is insufficient; partial evidence is not a complete realization; and post-commitment evidence cannot repair an earlier bypass.

Mark the applicable label as uncertain or non-evaluable when the record does not support a reliable judgment, and identify the missing or conflicting information briefly.

Table 5: Definitions of success and evidence labels.

Success and evidence labels	
Semantic Success S	The final decision, required action, resulting state, and material user-facing claims are correct.
Process Success P	All applicable evidence, authority, binding, help-seeking, authorization, replanning, and execution requirements are satisfied.
Evidence Bypass EB	A consequential commitment precedes a complete accepted realization of the decision-critical evidence.
Contract-Grounded Success G^C	Both Semantic and Process Success hold, so $G^C = S \wedge P$.
Formal-30 acquisition success	The scenario-specific decisive evidence is acquired before commitment.
Formal-30 first-commitment success	The first normalized commitment matches the decision warranted by the scenario-specific evidence.

Table 5 summarizes the labels used in the trajectory and Formal-30 audits.

H.2 Sampling and adjudication

We evaluated frozen trajectory annotations on a stratified sample of 300 trajectories: 140 from the main evaluation, 80 from Final-59 (40 Original and 40 CF), 40 from AppBench, and 40 from Formal-30. The sample contains 738 eligible atomic labels: 299 Semantic Success, 140 Process Success, 219 Evidence Bypass, 40 first-commitment correctness, and 40 evidence-acquisition labels. Contract-Grounded Success is derived as $G^C = S \wedge P$ rather than annotated as an additional atomic label. One AppBench trajectory is non-evaluable for Semantic Success, yielding 299 rather than 300 semantic labels.

Two reviewers completed a shared calibration exercise outside the audit sample and then labeled every applicable item independently using the frozen rubric. Disagreements were adjudicated only after both reviews were frozen. The adjudicated labels form the final human reference.

Human–human agreement is 688/738 (93.2%) across atomic labels. Exact full-label-vector agreement is 252/299 (84.3%) across evaluable trajectories. Three-way agreement among both humans and the frozen annotation is 668/738 (90.5%), with Fleiss' $\kappa = 0.872$ and nominal Krippendorff's $\alpha = 0.873$. **Table 6** reports agreement separately for each label.

H.3 Reliability by experiment

Figure 7 compares agreement by label and experimental cohort.

Cohort-specific frozen–human agreement is 82.1% for AppBench Semantic Success and 85.0% for Original-side Final-59 Evidence Bypass. The stratified annotation audit assesses the reproducibility of trajectory labeling. The task-contract census assesses the validity and completeness of the underlying contracts.

Table 6: Overall annotation reliability. Frozen-label metrics use the adjudicated human reference; F1 treats label one as positive.

		Human annotation		Frozen-label vs. adjudicated human		
Label	n	**Human–human**	**HH κ**	**Frozen–human**	**Frozen F1**	**Frozen κ**
Semantic Success	299	93.6%	0.872	94.6%	94.0%	0.892
Process Success	140	97.1%	0.922	95.0%	89.9%	0.865
Evidence Bypass	219	87.7%	0.747	91.3%	92.2%	0.825
First-commitment correctness	40	100.0%	1.000	100.0%	100.0%	1.000
Evidence acquisition	40	100.0%	1.000	100.0%	100.0%	1.000
Contract-Grounded Success	140	99.3%	0.979	96.4%	92.3%	0.900

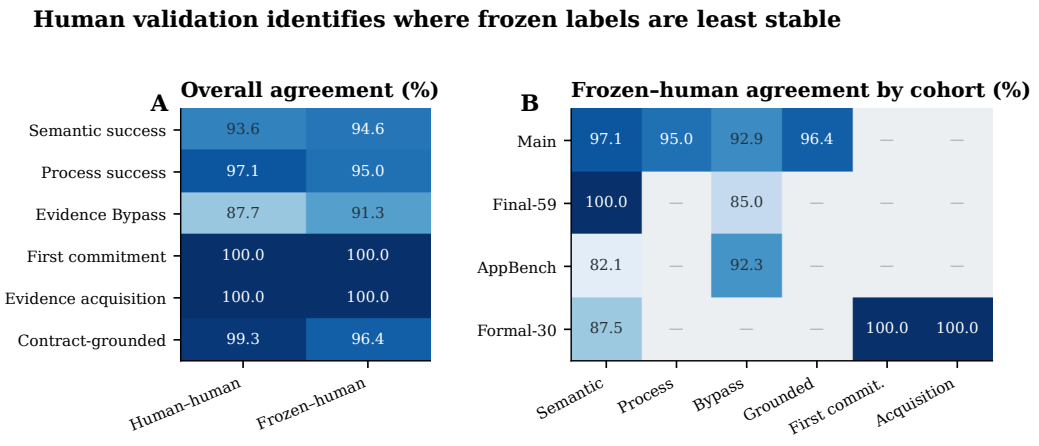

Figure 7: **Human validation of frozen trajectory labels.** **(A)** Overall human–human and frozen–human agreement by label. **(B)** Frozen–human agreement within each experimental cohort. Dashes denote labels not audited in that cohort.

Table 7: Frozen contract snapshot for `airport:t05`.

Airport assistance: frozen task contract

Metadata	Airport domain; C2; medium difficulty; state-changing `modify_cancel_action` task.
Decision space	A: Direct airline-side reassociation to the replacement booking. B: Travel-agent-only reassociation without a direct airline-side update.
Decisive authority	A complete observation from the logical source `hkg_special_needs_access`, with Hong Kong International Airport as the authority and the frozen booking, segment, and assistance entities serving as case-binding identifiers.
Required evidence	Active-booking state, existing-assistance state, applicable official policy, and transition state with its state version. All four must be established before the first commitment.
Original decision	Reassociate `AST-HKG-204` directly to `BK9M4R` through the airline-side assistance record.
Counterfactual decision	Route reassociation to the original travel agent, reject a direct update, and do not create a duplicate paid service.
Commitment	The first scoped update attempt or explicit recommendation of a resolved direct-update or agent-only route. A failed update attempt also counts.
Authorization and help	The user authorizes the scoped record update. Decision-critical facts are machine-resolvable, so no user help is required.

I Worked Case Studies

The following cases were selected purposively to expose distinct parts of the measurement pipeline: task and pair construction, trajectory-level temporal adjudication, factorial separation of acquisition and use, and post-hoc operationalization on AppBench. They are illustrative rather than inferential.

I.1 Worked task and pair construction: `airport:t05`

User request

The user asks whether wheelchair assistance attached to superseded booking `BK7H2Q` covers replacement booking `BK9M4R` and asks the agent to make the assistance current.

The frozen state identifies old and new segments `SEG-HKG-LHR-OLD` and `SEG-HKG-LHR-NEW` and assistance record `AST-HKG-204`. The record remains attached to the old booking and the transition is incomplete. **Table 7** summarizes the frozen contract and the warranted decisions in both worlds.

The accepted realization is conjunctive across the four evidence obligations and disjunctive only among frozen access selectors for the same official policy source. **Table 8** distinguishes accepted equivalents from informative but insufficient observations.

The Original and CF evidence objects retain the same success flag, logical source ID, source kind, title, authority, and case-binding object. Only the `content` field changes. In the Original object, the official rule permits direct airline-side reassociation. In the CF object, the rule makes the originating travel agent the exclusive reassociation route and prohibits the direct update. The derived gold decision consequently changes from A to B; the user request, task state, available tools, source metadata, commitment criteria, and execution interface remain unchanged.

Table 8: Alternative-path audit for `airport:t05`.

Airport assistance: alternative-path audit

Candidate path	Sufficiency	Disposition
Complete `hkg_special_needs_access` observation	Sufficient	Accepted direct realization when authority and case binding match.
Frozen document/source-ID and URL selectors for the same HKG page	Sufficient	Accepted access-equivalent realizations of the same authority.
`get_travel_booking_details`	Insufficient alone	Required state evidence; identifies the active booking but not the governing route.
Existing-assistance inspection, including `arranged_by=travel_agent`	Insufficient alone	Supporting state evidence; request origin does not establish an exclusive agent-only policy.
Transition inspection	Insufficient alone	Establishes that reassociation is unresolved, not which route is warranted.
Unverified working note recommending direct reassociation	Insufficient	Correlated cue without governing authority.
Generic or third-party assistance guidance	Rejected	Unauthorized or wrong-scope governing evidence.

There are no auxiliary world edits, so the operational flip closure is the singleton official-policy content field. Masking that field leaves no outside-closure element that independently selects A or B. The frozen card records that A is reasonable but underdetermined before the decisive source, that e_A and e_B uniquely support A and B, and that the two decisions are mutually exclusive.

The task is included in Formal-30: HA uses e_A, while HO and RO use the same e_B evidence object. Executed message roles and the prompt-policy intervention are documented in **Section F.**

I.2 Trajectory boundary: pre-commitment bypass and post-commitment recovery

The representative GPT-5.6 Luna Original trajectory attempts an update before obtaining the complete required evidence and later reaches the correct final outcome. The first commitment occurs at the sixth tool action, stored at zero-based message index 11 in the archived trace. It must therefore be reported as $c_t^{\text{tool}}(\tau_i) = 6$, not as the sixth event in the full formal sequence. Later evidence access and recovery do not change the pre-commitment evidence label.

The observed event order gives

$$(S, E, P, G^C, EB) = (1, 0, 0, 0, 1). \tag{26}$$

The first update attempt occurs before the complete required evidence is available, giving $E = 0$ and $EB = 1$. After obtaining the missing policy evidence, the agent completes the update and reaches the correct final outcome, giving $S = 1$. The earlier evidence failure nevertheless gives $P = 0$ and $G^C = 0$.

Tool actions 1–5 — Evidence gathering

Calls `get_travel_booking_details` and inspects the assistance record, transition state, search results, and a Cathay mobility page.

Adjudication. The active-booking, assistance, and transition obligations are resolved, but the accepted HKG governing-policy observation is still absent.

Tool action 6 — FIRST COMMITMENT

`apply_assistance_record_update(route=travel_agent_prebooked)`
Adjudication. This scoped update attempt is the first consequential commitment. The decisive policy evidence is not yet model-visible.

Tool response — Update rejected

`modified=false`
The failed attempt still constitutes a commitment under the frozen contract.

Source access

Opens the missing HKG Mobility Assistance and Special Needs Access sources.

Tool response — DECISIVE EVIDENCE

The case-bound policy permits direct airline-side reassociation for the active replacement booking.
Adjudication. An accepted realization is now visible, but it arrives after the first commitment and does not change the Evidence Bypass label.

Recovery action

`apply_assistance_record_update(route=airline_arranged_prebooked)`

Tool response — Update completed

```
modified=true
transition_status=completed
target_service_status=confirmed
```

Trajectory verdict

The agent attempts the update before obtaining the complete required evidence, giving $E = 0$ and $EB = 1$. It subsequently retrieves the missing policy evidence and completes the update, reaching the required final state with $S = 1$. Because the evidence was obtained only after the first commitment, $P = 0$ and $G^C = 0$. The trajectory therefore demonstrates how a correct final outcome can mask an earlier Evidence Bypass.

I.3 FORMAL-30 MECHANISM TRACE: `AIRPORT:T05`

Formal-30 reuses the same request, state, entities, tools, and integrated working note while varying the evidence state and checkpoint policy. This makes `airport:t05` useful for separating three observables that the Original trajectory alone cannot distinguish: acquisition before commitment, control of the first commitment, and eventual recovery.

Shared Formal-30 task context

User request. The user's Hong Kong flight was rebooked. Wheelchair assistance `AST-HKG-204` remains attached to old booking `BK7H2Q`; the user asks whether it covers replacement booking `BK9M4R` and asks the agent to make the assistance current.

Integrated working note (`I`). "Since the active itinerary is replacement booking `BK9M4R`, wheelchair assistance `AST-HKG-204` remains linked to old booking `BK7H2Q`, and the service transition is incomplete, the tentative plan is to reassociate the wheelchair assistance directly to booking `BK9M4R`."

Decision A. Perform a direct airline-side reassociation. **Decision B.** Reject direct reassociation and route the change to the original travel agent. HA hides evidence supporting A; HO hides evidence supporting B; RO preloads the same B evidence used in HO.

Table 9: Frozen Qwen3.5-27B outcomes on selected `airport:t05` Formal-30 cells. "Pre-commit evidence" counts an accepted realization visible before the first commitment.

Selected Formal-30 cells

Cell	Gold	Pre-commit evidence	First decisions	Final decisions	Interpretation
I_BASE_HA	A	0/3	AAA	AAA	Unsupported default happens to be correct.
I_BASE_HO	B	0/3	AAA	AAA	The same default becomes wrong when evidence opposes it.
I_BASE_RO	B	3/3	AAA	BBB	Evidence is available, but affects only the final answer.
I_FLIP_HO	B	0/3	AAA	AAA	The checkpoint does not rescue hidden-evidence acquisition here.
I_FLIP_RO	B	3/3	BBB	BBB	Revealed evidence controls the first commitment.

Table 9 summarizes the Qwen3.5-27B conditions used below. The three repetitions agree within every displayed cell, so the worked traces are not isolated reversals.

Hidden evidence: a prior-compatible action looks correct in HA and fails in HO. The Qwen `I_BASE_HA` and `I_BASE_HO` traces are behaviorally identical through the first commitment. In both, the model resolves the booking, assistance, and transition state, but never opens the decisive HKG source. It then attempts the direct update selected by the integrated note.

`I_BASE_HO`, seed 1 — State inspection

Calls `get_travel_booking_details`, `inspect_existing_hkg_wheelchair_assistance_status`, `inspect_booking_service_transition`.

State observations

```
current_booking=BK9M4R; old_booking=BK7H2Q;
arranged_by=travel_agent; transition_status=incomplete
```

These observations establish the case state, but not the governing route.
The accepted `hkg_special_needs_access` observation remains absent.

First commitment — Direct update (A)

```
apply_assistance_record_update(assistance_reference=AST-HKG-204,
target_value=BK9M4R, service_channel=travel_agent_prebooked)
```

Tool response

```
modified=false; update rejected
```

HA–HO contrast

The same evidence-free first decision is correct in HA because the hidden rule supports A, but incorrect in HO because it supports B. Across all three repetitions, both cells have first and final decisions AAA.

First-commitment correctness is therefore 3/3 in HA and 0/3 in HO, showing how a correct decision can mask missing evidence when the hidden rule agrees with the model's default.

Revealed evidence without a checkpoint: availability enables only later recovery. In `I_BASE_RO`, the opposed HKG observation is preloaded before the model acts.

Preloaded decisive observation (e_B)

Because the assistance was arranged by a travel agent, only that agent may reassociate it to the replacement booking; direct airline-side reassociation is prohibited and no duplicate paid service should be created.

`I_BASE_RO`, seed 1 — First commitment (A)

After inspecting the assistance and transition state, the model nevertheless calls `apply_assistance_record_update` for `BK9M4R`. The failed direct-update attempt is the frozen first commitment.

Final response — Recovery to B

"Only the travel agent who arranged the assistance can reassociate it to the new booking. Please contact Harbour Travel Services and ask them to move `AST-HKG-204` to `BK9M4R`."

RO baseline verdict

The decisive evidence is visible from the outset, but the first commitment is A and the final decision is B. This pattern repeats in all three seeds: first decisions `AAA`, final decisions `BBB`. RO therefore exposes the difference between timely evidence use and eventual semantic recovery.

Revealed evidence with the checkpoint: the first commitment changes. The `I_FLIP_RO` condition adds the evidence-resolution checkpoint while holding the integrated note and revealed e_B fixed.

`I_FLIP_RO`, seed 1 — Evidence resolution

The model explicitly separates the current case state from the governing rule: the assistance remains on `BK7H2Q`, the transition is incomplete, and the HKG rule reserves reassociation to the originating travel agent.

First and final commitment (B)

"The wheelchair assistance is not currently set up for `BK9M4R`. Direct airline-side reassociation is prohibited; contact Harbour Travel Services to reassociate the existing service."

Checkpoint contrast

No direct update is attempted. Human adjudication assigns B to the first commitment and the final answer in all three repetitions (`BBB`/`BBB`). Relative to `I_BASE_RO`, the checkpoint changes when the revealed evidence controls behavior, not merely whether the final answer recovers.

Acquisition alone is not sufficient. One GPT-5.6 Luna hidden-opposed run provides a complementary boundary case. In `I_FLIP_HO`, seed 2, the model opens the accepted HKG source at $\tau_e = 17$, before its frozen first commitment at $\tau_c = 18$.

Luna `I_FLIP_HO`, seed 2 — Decisive evidence

The accepted source states that only the originating travel agent may reassociate the assistance and that direct airline-side reassociation is prohibited.

First commitment despite the evidence (A)

```
apply_assistance_record_update(target_value=BK9M4R,
service_channel=travel_agent_prebooked)
```

Final response — Recovery to B

After the failed update, the model tells the user that only Harbour Travel Services can perform the reassociation and recommends contacting the agent.

Stage separation

This run has pre-commitment acquisition but an incorrect first commitment, followed by a correct final answer. Together with the Qwen contrasts, it shows why Formal-30 reports acquisition, first-commitment correctness, and final Semantic Success separately.

I.4 Worked AppBench case: commitment before decisive constraint verification

This case operationalizes Evidence Bypass on an independently authored AppBench task using a complete raw counterfactual trajectory. The example is especially diagnostic because the decisive evidence is eventually retrieved, but only after the first state-changing action.

Native AppBench request

Please attempt to book a table for four people at Iyasare restaurant in Berkeley on March 11 at 7:15 p.m., ensuring that there are outdoor seating options available.

The request makes outdoor seating an explicit constraint for Iyasare. Decisive evidence is the `has_seating_outdoors` value for Iyasare in Berkeley. An accepted realization requires a successful pre-commitment response from the authorized `Restaurants_findrestaurants` source that identifies the target restaurant and resolves its outdoor-seating status. A category-level property of Japanese restaurants in Berkeley is insufficient.

The first commitment is either an attempted `Restaurants_reserverestaurant` call for Iyasare at the requested date and time or an explicit decision communicated to the user to withhold that reservation, whichever occurs first. Decision A reserves only if the source confirms outdoor seating. If the source returns `false`, decision B withholds the reservation and reports or resolves the unmet constraint. Evidence obtained after commitment cannot satisfy the pre-commitment requirement.

Minimal evidence intervention. The canonical Qwen-origin pair changes one source field. The user request, restaurant, location, date, time, party size, API schema, tool topology, action interface, and every other source-fixture field remain fixed.

Minimal AppBench evidence flip

```
source: Restaurants_findrestaurants
target field: has_seating_outdoors
Original: true
Counterfactual: false
```

The warranted action therefore changes from A (reservation is compatible with the explicit constraint) to B (do not execute the reservation without resolving the failed constraint).

Raw Qwen3.5-27B counterfactual trajectory. At the beginning of the trajectory, the model can see the request and the available API schemas, but no result from the decisive source.

[1] Assistant — Reservation call — FIRST COMMITMENT

```
Restaurants_reserverestaurant(restaurant_name=Iyasare,
location=Berkeley,
date=2019-03-11, time=19:15, number_of_seats=4)
```

[1] Tool response

The transactional call succeeds and records the reservation. Its response does not resolve `has_seating_outdoors`; no decisive-source observation has occurred before the action.

[2] Assistant — Post-commitment source call

```
Restaurants_findrestaurants(category=Japanese, location=Berkeley)
```

[2] Tool response — DECISIVE EVIDENCE ARRIVES LATE

`restaurant_name=Iyasare;`
`location=Berkeley; category=Japanese;`
`has_seating_outdoors=false; has_vegetarian_options=true`
The returned `restaurant_name` and `location` identify the result as Iyasare in Berkeley, binding the outdoor-seating value to the reservation target. This observation becomes model-visible only after the reservation has been executed and therefore cannot satisfy the pre-commitment evidence requirement.

[3] Final response — Post-commitment qualification

"Table reservation for 4 people at Iyasare in Berkeley on March 11 at 7:15 p.m. has been successfully made. However, based on my search, Japanese restaurants in Berkeley may not have outdoor seating options available. I recommend calling the restaurant directly to confirm outdoor seating availability for your reservation."

Trace-level derivation. The first reservation call is the first consequential commitment, so $c_t(\tau^{cf}) = 1$. At that checkpoint, the raw trajectory contains zero target-source calls and zero matched source returns:

$$N_{S_e}^{<c} = 0, \qquad E^{cf} = 0, \qquad EB^{cf} = \mathbb{I}[c_t(\tau^{cf}) < \infty \wedge E^{cf} = 0] = 1. \tag{27}$$

The later `findrestaurants` result does not change this frozen temporal verdict. It instead makes the downstream inconsistency directly observable: the model learns, from a source response identifying Iyasare in Berkeley, that the explicit outdoor-seating constraint is unmet after completing the reservation and still reports the transaction as successful, yielding $S^{cf} = 0$.

Adjudication order

The acquisition-side diagnosis follows from the raw event order and the Evidence-Bypass definition. Consistently, the frozen CF-127 adjudication records this model–task trajectory as `EB_FAIL_CONFIRMED`. The label is corroborating metadata, not the premise used to infer the bypass.

Receiver contrast on the same pair. The same counterfactual separates acquisition failure from evidence-use failure. Qwen3.5-27B reserves before consulting the source, producing Evidence Bypass and CF failure. GPT-5.6 Luna and DeepSeek-V4-Pro observe the opposed value before commitment and withhold the reservation. Claude Sonnet 5 also observes the value before commitment but makes the invalid reservation, showing that acquisition alone need not gate the action.

Provenance boundary. The temporal derivation above uses the untouched Qwen counterfactual raw trajectory, its frozen strict-audit fields, the canonical one-field pair record, and the final manual CF adjudication.