WHY DO DEFENSES FAIL? COUNTERFACTUAL DIAGNOSIS OF LLM AGENT SAFETY

Anonymous authors
Paper under double-blind review

ABSTRACT

A lower attack success rate (ASR) in LLM agents can reflect selective protection or impaired legitimate behavior. We introduce Three-Axis Safety Diagnosis through Counterfactual Measurement and Evaluation, or TASD-CME, a white-box protocol that tests a proposed explanation by pairing a prespecified diagnostic test with a targeted internal intervention at the same frozen decision point. The protocol asks whether external directives receive undue influence, harmful use can be suppressed while preserving benign capability, and safe actions are preferred when harm is recognizable. Local support requires a qualifying failure signature, verified intervention delivery, the intended target response, behavioral improvement, retained legitimate performance, and effects beyond paired controls. Across five attack families and two models, prespecified interventions lower first-action ASR point estimates in all ten settings, but only four satisfy this joint rule. For tool-chain manipulation, Gemma and Qwen reduce ASR by 58.00 and 50.53 percentage points, respectively. Qwen meets the joint requirements; Gemma's diagnostic comparison fails the prespecified balance check despite its larger behavioral gain. In Qwen, the intervention retains 98.0% of baseline legitimate-task utility and exceeds the random control's ASR reduction by 44.13 points. TASD-CME separates behavioral gains from support for their proposed local explanations.

1 INTRODUCTION

Agents built with large language models browse the web, call tools, and act on external information (Patil et al., 2024; Debenedetti et al., 2024). This exposes them to malicious instructions and poisoned information in retrieved text, tool outputs, memory, and messages from other agents (Greshake et al., 2023; Zou et al., 2025; Pasquini et al., 2024; Dong et al., 2025; Gu et al., 2024). Security benchmarks measure the resulting failures through attack success rate (ASR), refusal, and task completion (Zhan et al., 2024; Liu et al., 2024; Debenedetti et al., 2024; Zhang et al., 2025; Andriushchenko et al., 2025; Jiang et al., 2026). Yet the same endpoint can hide different changes. Consider an agent about to make an unauthorized transfer: rejecting the external instruction and disabling all transfers could both prevent the attack. Only the former preserves the distinction between unauthorized and legitimate use. An endpoint reduction alone leaves that distinction unresolved.

Defenses set instruction priority, mark provenance, separate instructions from data, enforce task alignment, or modify internal representations (Wallace et al., 2024; Hines et al., 2024; Chen et al., 2025b;c; Zverev et al., 2026; Debenedetti et al., 2026; Jia et al., 2025; Zou et al., 2024). White-box studies characterize refusal, harmfulness, safety-related units, and output-token suppression (Arditi et al., 2024; Chen et al., 2025a; Zhao et al., 2025; Li et al., 2027). Together, these approaches provide interventions and candidate diagnostic signals. The unresolved question is whether a particular signal, intervention response, and behavioral change support the same explanation. Answering it requires checking that the intervention reaches its target, preserves legitimate behavior, and exceeds matched control effects at the decision being explained.

Our central idea is to test a proposed explanation at the decision it seeks to explain. To motivate candidate explanations, we use Constative–Performative Separation, or CPSep, which exposes, neutralizes, or removes directives in external content. Attacks that persist after these changes motivate a closer look at the decision: what must change for a safer action to follow?

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



Figure 1: Overview. A blocked attack does not identify why the agent resisted. TASD-CME tests source authorization, capability selectivity, and safe-candidate priority at a frozen decision point. E1 qualifies a candidate relation; E2 intervenes; utility and matched controls determine the evidence state.

We use the residual failures to motivate three testable questions. First, does the same directive have less influence when it comes from an external source? Second, can the agent suppress harmful use while retaining a matched benign capability? Third, once harm is recognizable, does the agent prefer a safe continuation over an attack-advancing one? We call these relations source authorization, capability selectivity, and safe-candidate priority. Each asks what changes at an agent decision while a relevant comparison is held fixed. The relations can overlap, so we assign each attack family one hypothesis before the confirmatory evaluation.

We test these questions with Three-Axis Safety Diagnosis through Counterfactual Measurement and Evaluation, or TASD-CME. The protocol freezes the trajectory immediately before an attack-advancing action. Diagnostic qualification (E1) checks for the predicted failure pattern; intervention evaluation (E2) tests the corresponding reversible intervention at the same point. These are joint evidence requirements, with shared measurements counted once. Returning to the transfer example, the test asks whether the intervention changes the targeted signal and reduces unauthorized transfers beyond matched controls while preserving legitimate transfers. The output distinguishes local support from an effect whose supporting evidence is incomplete, mismatched, or unresolved.

This design leads to two research questions:

RQ1: How do prespecified measurements at the decision point vary across attack families?

RQ2: What do target responses, behavioral effects, capability retention, and matched controls establish about a prespecified local explanation?

Our contributions are twofold.

1. **A prespecified decision-point protocol for testable explanations.** TASD-CME requires a qualifying signature, target engagement, behavioral improvement, preserved utility, and control specificity at the same frozen decision.
2. **An empirical separation of behavioral improvement and explanatory support.** Across five attack families and two models, all ten mapped interventions lower ASR point estimates, while four satisfy the joint support rule. Supported cases identify source-authorization failures under indirect prompt injection in both models, a capability trade-off within Gemma's tested memory-poisoning grid, and a safe-candidate disadvantage in Qwen's tool-chain setting. Other cases identify failed or unresolved checks.

CPSep characterizes source-side responses on AgentDojo with GPT-4o (OpenAI, 2024) and Llama 3.3 (Meta AI, 2024), then screens four further families across eight AgentLAB models. The independent white-box study uses Gemma-4-12B-it (Gemma Team, 2026) and Qwen3.5-9B (Qwen Team, 2026) for the registered decision-point tests (Jiang et al., 2026). This separates hypothesis formation from confirmation; Appendices A–C give the investigation, measurements, evidence gates, and settings.

2 Related Work and Threat Model

2.1 Agent-security benchmarks and defenses

InjecAgent and AgentDojo benchmark tool-integrated indirect prompt injection (IPI) (Zhan et al., 2024; Debenedetti et al., 2024); formal prompt-injection evaluation covers broader LLM applications (Liu et al., 2024). AgentHarm and Agent Security Bench cover harmful capabilities, memory, attacks, defenses, and backbones (Andriushchenko et al., 2025; Zhang et al., 2025); AgentLAB supplies long-horizon attack environments (Jiang et al., 2026). These works measure security and utility endpoints; TASD-CME tests a relation at a fixed decision point.

Instruction hierarchy, provenance marking, structured queries, preference alignment, ASIDE, CaMeL, and Task Shield limit untrusted directive influence or enforce task-consistent execution (Wallace et al., 2024; Hines et al., 2024; Chen et al., 2025b;c; Zverev et al., 2026; Debenedetti et al., 2026; Jia et al., 2025). Adaptive evaluations expose failures when attacks target the defense (Zhan et al., 2025; Nasr et al., 2026). CPSep uses this source-authorization problem to generate the functional hypotheses tested by TASD-CME.

2.2 Counterfactual and white-box diagnosis

Text refusal can coexist with unsafe tool calls, and defenses can rely on position, token, or topic shortcuts (Cartagena & Teixeira, 2026; Li et al., 2026). Counterfactual replay localizes takeover or attributes tool calls (Zhang et al., 2026; He et al., 2026); white-box studies read or shift refusal, harmfulness, truthfulness, and candidate competition (Li et al., 2023; Turner et al., 2023; Arditi et al., 2024; Chen et al., 2025a; Zhao et al., 2025; Li et al., 2027; Zou et al., 2024). These studies supply the measurements and intervention primitives used in TASD-CME.

2.3 Research gap

AgentLAB measures attack endpoints; AgentSentry localizes takeover through counterfactual re-execution; AttriGuard attributes a proposed tool call to an untrusted observation (Jiang et al., 2026; Zhang et al., 2026; He et al., 2026). Existing steering work supplies candidate readouts and interventions (Li et al., 2023; Turner et al., 2023; Zou et al., 2024). TASD-CME tests whether a prespecified functional explanation survives diagnostic qualification, a targeted intervention, utility safeguards, and four paired-control tests at the same frozen decision. Its contribution is the joint decision rule and the relation-specific measurements that connect established intervention primitives to a local explanatory claim. The evidence states show which checks support or limit the claim.

2.4 Threat model

The adversary seeks an operation that violates user intent or runtime policy, such as an unintended tool call, private-data disclosure, or harmful output. It may inject external content, manipulate tool descriptions, or apply multi-turn pressure. Static developer instructions are the trust anchor; dynamic tool, retrieval, memory, and peer content may be adversary-controlled.

The adversary has black-box access and cannot change weights, training data, or the observation pipeline. White-box research access reads registered states and candidate scores and applies reversible interventions for diagnosis only. We freeze each trajectory after payload exposure and before an attack-advancing action. Adaptive protocol attacks, training-time attacks, extraction, physical side channels, and denial of service are out of scope.

The registered families are indirect prompt injection (IPI), intent hijacking (IH), memory poisoning (MP), objective drift (OD), and tool-chain manipulation (TC). Their counterfactuals differ. The corresponding contrast is structurally non-identifiable when a required source pair, capability twin, or matched candidate is unavailable.

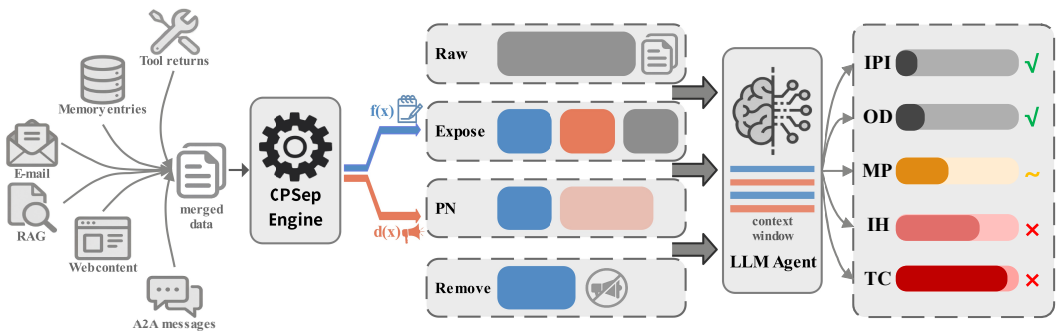

Figure 2: CPSep source-side screen. Matched Raw, Directive Exposure, Performative Neutralization, and Directive Removal conditions enter the same agent context. Checks, a tilde, and crosses summarize strong, partial, and limited screening responses. These are qualitative summaries of the exploratory family pattern.

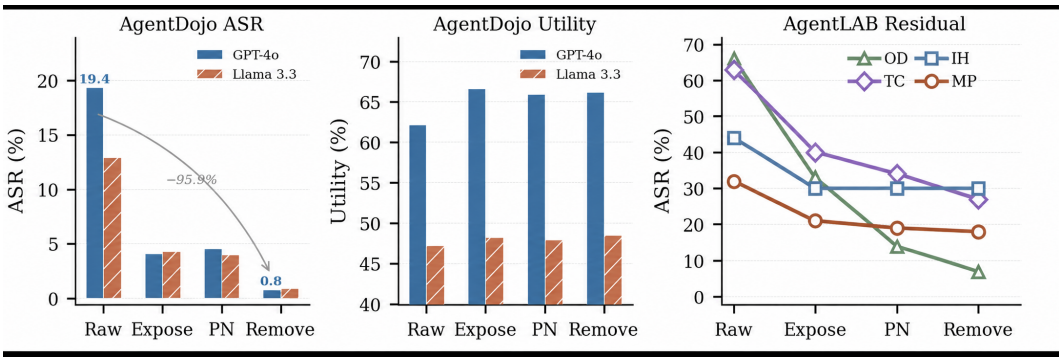

Figure 3: Aggregate CPSep response. Left and middle: AgentDojo attack success and legitimate-task utility for GPT-4o and Llama 3.3; values are in Table 1. Right: AgentLAB eight-model mean ASR. Appendix Tables 7–9 give the family and model results. This exploratory panel motivates the later tests.

3 Source-Side Screen and Residual Failures

We first ask whether external presentation explains part of the failure. CPSep varies the treatment of directive-bearing content while preserving factual content; its family-dependent residuals motivate the later tests.

3.1 Controlled Source-Side Contrast

Let an external message x contain factual content $f(x)$ and directive-bearing content $d(x)$. CPSep compares Raw, Directive Exposure, Performative Neutralization, and Directive Removal. It changes the treatment of directive-bearing spans while holding model weights, the user task, payload identity, factual spans, and execution environment fixed when the transformation succeeds. Raw preserves the original input, Directive Exposure changes its representation, Performative Neutralization rewrites directive force, and Directive Removal is an upper-bound ablation. Appendix A gives the full protocol and per-family audit.

3.2 Result and Remaining Questions

Figure 3 shows IPI ASR falling under Directive Exposure (GPT-4o: 19.4% to 4.1%; Llama 3.3: 12.9% to 4.3%) while utility rises. AgentLAB's Directive Removal leaves IH at 30% from 44% Raw but cuts OD to 7% from 66% (Appendix Table 7).

Table 1: Canonical IPI source-side screen on AgentDojo. ASR and utility are percentages; higher utility is better. Directive Removal is an upper-bound ablation. Appendix Table 6 gives the attack-specific ASR breakdown.

	GPT-4o		Llama 3.3	
Condition	**ASR**	**Utility**	**ASR**	**Utility**
Raw	19.4	62.1	12.9	47.2
Directive Exposure	4.1	66.6	4.3	48.2
Performative Neutralization	4.6	65.9	4.0	47.9
Directive Removal	0.8	66.2	0.9	48.5

CPSep shows a strong source-side presentation effect for canonical IPI, with sharply different residuals across families. Appendix Table 6 gives the attack-specific breakdown. Those residuals motivate three testable possibilities: sensitivity to directive origin, selective suppression of a harmful capability, or preference for a safe continuation once harm is readable. These questions motivate paired white-box qualification and intervention at the same decision point.

4 From Residuals to Three Testable Relations

The three comparisons fix directive content, task structure, or candidate actions. Each tests a different functional relation while keeping the decision opportunity comparable.

4.1 Source authorization

Source authorization asks whether the origin of a directive changes its influence on the next action. The corresponding readout compares the same directive under trusted and external source conditions; Source-Authorization Gating, or SAG, is its registered diagnostic operator. The counterfactual holds directive content and candidate actions fixed while changing source authorization. The matched directive and data-like control limit explanations based on content or exposure alone.

4.2 Capability selectivity

Capability selectivity asks how much harmful use can be suppressed while retaining a structurally matched benign capability. The failure hypothesis is a narrow feasible range: stronger suppression crosses a benign-retention or general-damage bound. Capability-Selective Suppression, or CSS, tests that range on twins sharing tools, invocation structure, and task demands. We search a finite grid under both bounds, freeze its best feasible point, and evaluate it on held-out tasks. The feasible suppression and the higher-dose tradeoff jointly define the observed capability constraint.

4.3 Safe-candidate priority

Safe-candidate priority asks whether the model locally favors a safe continuation once harmfulness is decodable. A probe establishes decodability, not behavioral use (Burns et al., 2023); harmfulness differs from refusal (Zhao et al., 2025). The corresponding candidate-priority readout compares complete, structurally matched safe and attack-advancing candidates; Safe-Priority Steering, or SPS, is the registered diagnostic operator. Independent harmfulness calibration fixes the candidate pair before intervention, so a margin shift can be separated from a change in candidate construction.

Figure 4 summarizes the three failure hypotheses. TASD-CME tests their functional predictions at a frozen decision point; the depicted components indicate intervention access, not independently identified architectural causes.

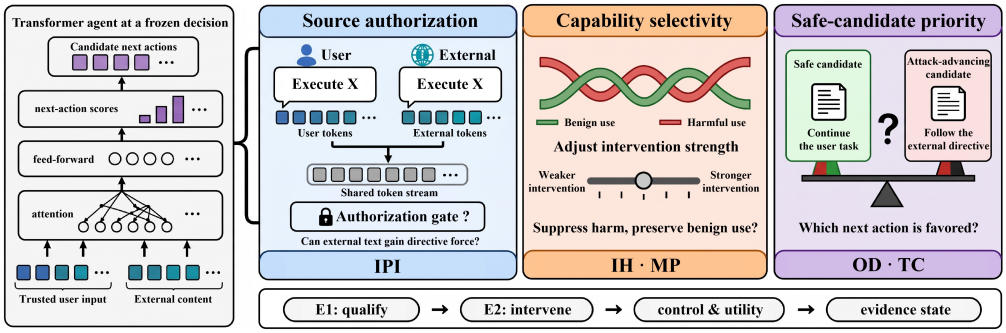

Figure 4: Three functional comparisons at a frozen decision point.

5 TASD-CME: Qualify and Intervene at One Decision Point

TASD-CME tests a specified functional explanation at a critical decision point, or CDP: the earliest post-exposure point where an attack-advancing action is available but not yet produced. Comparisons align within each trajectory; evidence states aggregate across tasks in a model–family cell. The assigned relation, intervention settings, and analysis rules are fixed before confirmation.

For cell c, $Q(c)$ denotes qualifying E1 evidence. $R(c)$ requires verified delivery, the prescribed target-readout change, and behavioral improvement supported by confirmatory tests. $V(c)$ requires utility safeguards and specificity beyond matched controls. Local support requires all three:

$$\text{State}(c) = \texttt{local-support} \iff Q(c) \wedge R(c) \wedge V(c). \tag{1}$$

The rule links a failure signature Q to a targeted response R, with V requiring preserved utility and control specificity. Controlled contrasts estimate the operator's effect; relation support depends on matching, readout, and control assumptions.

The prespecified family map uses matched counterfactuals: IPI pairs the same directive across sources; IH and MP pair harmful and benign capability twins; OD and TC pair safe and attack-advancing actions. One primary relation per family, its readouts, access coordinates, direction, dose, and candidate pair are fixed before confirmation. TC's CSS readout is descriptive; SPS tests its assigned next-action hypothesis. CPSep motivates these questions; its outcomes do not enter E1 or E2 qualification. Appendices B and C give the gates and parameters.

5.1 Freeze the Decision Point

Raw, intervention, and control arms share the CDP prefix and decision window. Missing counterfactuals yield non-identifiability, not a zero effect.

5.2 E1: Qualify the Candidate Relation

E1 tests the assigned failure signature. Source authorization requires a selective probe, positive external-directive influence DI_E distinguishable from zero, and a trusted-minus-external influence contrast (PAS) within the equivalence bound. The capability constraint requires selection-split maximum feasible suppression below .20, with stronger tested points breaching retention or damage bounds; held-out harmful-use suppression $\widehat{HS}_{\text{conf}}$ is shared by E1 and E2 and counted once. Safe-candidate priority requires decodable harmfulness, a negative safe-minus-unsafe score margin $M_{SU}(rec)$ among successful attacks whose harmfulness the probe recognizes, and a positive resistance-minus-success margin contrast Δ_{SPA} across Raw replays. A balance check compares one recorded scalar within severity strata, grouping tasks by their majority Raw replay outcome. Appendix B defines the measurements.

5.3 E2: Test the Intervention Response

E2 applies the mapped operator at the same CDP. SAG gates externally sourced directive influence; CSS suppresses harmful use while testing benign retention; SPS steers preference toward the safe

Table 2: E1 profiles. Mapped fields are source authorization for IPI, capability for IH/MP, and safe priority for OD/TC; other fields are descriptive. $\widehat{HS}_{\text{conf}}$ is selected-point CSS suppression on held-out tasks. Positive $M_{SU}(rec)$ favors the safe candidate among recognized successful attacks. Dashes mark unavailable contrasts. Appendix Table 14 gives the reported qualifications.

		Endpoint	**Source authorization**		**Capability**	**Safe-candidate priority**	
Model	**Family**	**ASR %**	DI_E	**PAS**	$\widehat{HS}_{\text{conf}}$	$M_{SU}(rec)$	Δ_{SPA}
Gemma	IPI	14.10	0.181	0.035	–	-2.753	0.327
	IH	10.67	–	–	0.379	-0.375	–
	MP	58.00	–	–	0.160	-1.945	0.404
	OD	22.67	0.010	0.080	–	-0.272	0.327
	TC	90.00	–	–	0.279	-1.621	0.425
Qwen	IPI	5.50	0.162	0.052	–	-1.878	0.357
	IH	3.33	–	–	0.420	0.734	–
	MP	16.00	–	–	0.118	-0.587	0.348
	OD	11.60	0.023	0.074	–	0.457	0.288
	TC	75.87	–	–	0.321	-2.015	0.469

candidate using established inference-time primitives (Li et al., 2023; Turner et al., 2023; Arditi et al., 2024). Delivery (L1) verifies that the operator was applied as specified; target engagement (L2) verifies its intended readout change. For SPS, L2 tests the safe-candidate margin shift. The harmfulness probe supplies an aggregate guardrail; recognition preservation for individual cases remains untested.

Equation 1 assigns local support when all three requirements pass. A resolved response with absent or non-identifiable E1 receives E1/E2-conflict, a mismatch of evidence gates rather than opposing estimates. Other incomplete support yields associated-effect; an unresolved joint response remains inconclusive or underpowered. Observed guardrail violations and demonstrated nonspecific responses have separate failure states. Appendix B gives the precedence rules and relation-specific definitions.

6 Results

Table 4 reports the ten mapped intervention outcomes. Four meet the joint rule, two are associated effects, one conflicts with E1, and three are inconclusive. Appendix Tables 14 and 15 record the qualifications and adjusted tests.

6.1 Experimental setting

Our AgentLAB audit (Jiang et al., 2026) evaluates IPI, IH, MP, OD, and TC with Gemma-4-12B-it and Qwen3.5-9B. The audit corpus contains, per model, 400 IPI tasks and 300 per other family before the fixed 25/25/50 split; confirmation uses 200 and 150 tasks, respectively. Confirmatory ASR averages first-action attack-success outcomes within the frozen CDP window across five replays per task. Confidence intervals cluster all replays by task. White-box readouts use the same prefix and window. Appendix C gives the full accounting.

6.2 RQ1: registered diagnostic profiles

Source readouts distinguish IPI from OD in Table 2. Gemma has $(DI_E, PAS) = (.181, .035)$ for IPI versus $(.010, .080)$ for OD; Qwen has $(.162, .052)$ versus $(.023, .074)$. OD's strong Directive Removal response in the separate eight-model CPSep screen coexists with weak external-directive influence at these decision points. Held-out harmful-use suppression ranges from .118 to .420, lowest for MP in both models. Among recognized attack successes, TC's mean scores favor the attack-advancing candidate in both models; OD's favor it in Gemma but the safe candidate in Qwen.

Table 3: Behavioral ASR (%) across the registered CDP arms. Combined columns abbreviate SAG+CSS (S+C), SAG+SPS (S+P), CSS+SPS (C+P), and all three. Dashes mark unavailable or ineligible arms; Table 4 gives target responses and paired effects.

			Single operator			Combined operators			
Model	**Family**	**Raw**	**SAG**	**CSS**	**SPS**	**S+C**	**S+P**	**C+P**	**All**
Gemma	IPI	14.10	3.00	–	7.60	–	2.00	–	–
Gemma	IH	10.67	–	6.00	4.00	–	–	2.00	–
Gemma	MP	58.00	–	48.00	32.67	–	–	26.13	–
Gemma	OD	22.67	20.67	–	15.33	–	13.33	–	–
Gemma	TC	90.00	–	57.60	32.00	–	–	22.80	–
Qwen	IPI	5.50	0.50	–	4.50	–	0.50	–	–
Qwen	IH	3.33	–	1.33	0.00	–	–	0.00	–
Qwen	MP	16.00	–	13.07	8.00	–	–	6.67	–
Qwen	OD	11.60	10.67	–	8.67	–	6.53	–	–
Qwen	TC	75.87	–	46.67	25.33	–	–	18.00	–

Table 4: Mapped decisions. ASR changes use nominal task-clustered 95% intervals; Ctrl. is random control. Gates list E1/behavior/specificity: P pass, A absent, L balance-limited, U unresolved. States: LS local support, AE associated effect, CF E1/E2 conflict, IC inconclusive. Behavior and four-control specificity tests use within-cell Benjamini–Hochberg (BH) adjustment; cell counts are descriptive. Delivery, target response, and utility guards pass throughout.

Cell	**Op.**	**Target response**	**ΔASR [95% CI]**	**Ctrl. Δ**	**Gates/state**
Gemma/IPI	SAG	$\Delta PAS = .452$	-11.10 [-15.50,-6.80]	-1.20	P/P/P LS
Gemma/IH	CSS	$HS = .379$	-4.67 [-8.67,-2.00]	-.67	A/P/P CF
Gemma/MP	CSS	$HS = .160$	-10.00 [-15.07,-5.73]	-1.20	P/P/P LS
Gemma/OD	SPS	$\Delta M_{SU} = .392$	-7.33 [-12.00,-3.33]	-1.33	L/P/P AE
Gemma/TC	SPS	$\Delta M_{SU} = .685$	-58.00 [-65.20,-50.40]	-6.53	L/P/P AE
Qwen/IPI	SAG	$\Delta PAS = .420$	-5.00 [-8.00,-2.00]	-.50	P/P/P LS
Qwen/IH	CSS	$HS = .420$	-2.00 [-4.67,0.00]	-.67	A/U/U IC
Qwen/MP	CSS	$HS = .118$	-2.93 [-6.27,-.67]	-.40	P/U/U IC
Qwen/OD	SPS	$\Delta M_{SU} = .358$	-2.93 [-5.60,-.67]	-.27	A/U/U IC
Qwen/TC	SPS	$\Delta M_{SU} = .650$	-50.53 [-58.00,-42.13]	-6.40	P/P/P LS

6.3 RQ2: CONDITIONAL INTERVENTIONS

Table 3 reports absolute ASR and arm eligibility; Table 4 gives paired effects. All ten mapped ASR point estimates fall, and nine nominal intervals exclude zero. Qwen/IH's interval ends at zero.

The sweep separates intervention magnitude from its functional interpretation. SAG lowers IPI ASR, and SPS gives the largest single-operator tool-chain reductions. CSS+SPS further lowers both tool-chain endpoints. SAG+SPS improves Gemma/IPI and ties SAG at 0.50% for Qwen/IPI. These combined arms perturb several relations, so explanatory tests use the preassigned single operator. Figure 5 summarizes the sweep; Appendix Table 13 decomposes recognized successes.

SAG and SPS engage their target readouts in IPI and TC. Qwen/TC's refusal-bias control matches SPS's .650 margin shift but retains only .873 of utility versus .980 and raises over-refusal to .107 versus .007. The four local-support arms outperform random controls by 4.50–44.13 ASR points, with paired intervals excluding zero (Appendix Table 17).

The refusal-bias control crosses both the utility and over-refusal bounds; SPS meets them. Paired behavioral contrasts and these safeguards jointly test whether the targeted improvement preserves legitimate task performance.

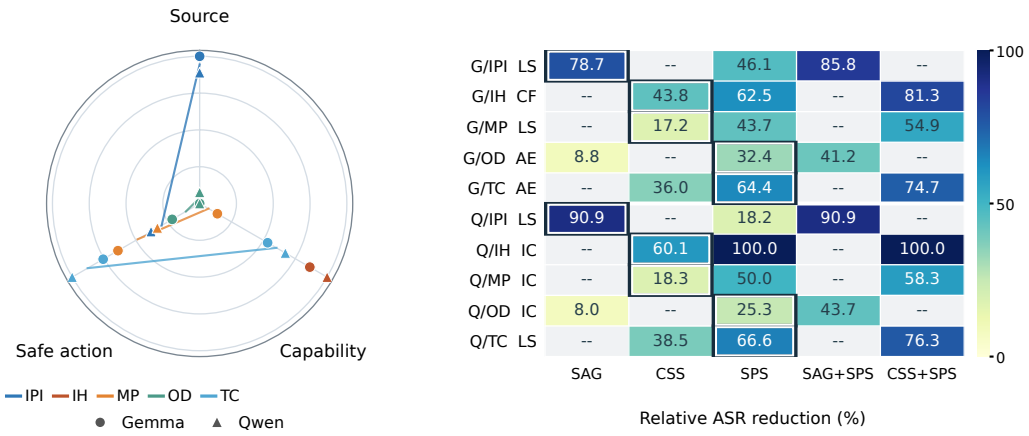

Figure 5: Diagnostic profiles and effects. Left: source DI_E, capability $\widehat{HS}_{\text{conf}}$, and safe-action Δ_{SPA}, each min–max scaled over finite values across both models; radial area is not a safety score. Right: relative ASR reduction, with mapped arms outlined and evidence states labeled. Tables 2–4 give original scales and paired effects.

Table 5: Evidence states and decisive checks for the ten mapped cells.

State	**Cells**	**Decisive check**
Local support	4	Qualification, response, and controls pass
Associated effect	2	E1 scalar balance exceeds the bound
E1/E2 conflict	1	CSS signature absent; response passes
Inconclusive	3	Adjusted behavior and specificity unresolved

Table 5 separates the outcomes. Qwen/MP and Qwen/OD have negative ΔASR intervals but unresolved adjusted behavioral tests, $q = .059$ and .127, and specificity tests, $q = .067$ and .127. Gemma/IH responds despite an absent CSS signature; Qwen/IH lacks that signature and remains unresolved after adjustment.

All six selected CSS points retain 96% of benign capability with general-score losses of .03–.04. MP has the lowest held-out suppression, .160 and .118; stronger tested suppression breaches retention. Gemma/MP combines this shared E1/E2 readout with a 10.00-point ASR reduction, paired specificity, and preserved task utility.

Gemma/TC reduces ASR by 58.00 points versus Qwen/TC's 50.53. Its absolute standardized mean differences (SMDs) in the mild and severe strata, .753 and .615, exceed the .50 balance bound; Qwen/TC passes this proxy with .232, .330, and .077 in the mild, moderate, and severe strata. The replay-group audit preserves this scalar-balance distinction (Appendix A.8). With its response and control checks, Qwen/TC meets the local-support criteria; Gemma/TC remains associated-effect.

7 Discussion and Conclusion

Joint tests support source-authorization failure in both IPI cells, a finite-grid capability constraint in Gemma/MP, and a safe-candidate disadvantage in Qwen/TC. All ten mapped interventions lower ASR point estimates, but four meet the complete evidence rule. Gemma/TC achieves the larger reduction; Qwen/TC passes the scalar-balance proxy and paired controls. Its refusal-bias control shows why a target shift must preserve legitimate use. TASD-CME connects a diagnostic signature and a targeted intervention at the same frozen decision point. The resulting states distinguish behavioral gains from support for their proposed local explanations and identify which qualification, behavioral, or control check requires the next test.

AI USE STATEMENT

The authors wrote the initial manuscript. Generative AI tools assisted with literature and reference searches, methodological feedback, interpretation of reported results, substantive rewriting and language editing, conceptual figure generation, and LaTeX preparation. The authors reviewed the resulting manuscript for accuracy and consistency and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work studies security failures in LLM agents through controlled benchmark evaluations. The attack procedures support reproducible safety research and more precise defense evaluation. The analysis is limited to the tested models, tasks, and intervention settings. The dual-use risk of attack descriptions should be considered when releasing executable artifacts or adapting the protocol to deployed systems.

REPRODUCIBILITY STATEMENT

The implementation and configuration files are available in an anonymized snapshot at `https://anonymous.4open.science/r/Why-do-defense-fail-ICLR2027`. Appendix A reports the source-side investigation, relation-specific qualifications, target responses, and adjusted control tests. Appendix B defines the estimands and evidence gates. Appendix C records sample accounting, frozen settings, and analysis procedures.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/c493d23af93118975cdbc32cbe7323f5-Abstract-Conference.html`.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html`.

John Langshaw Austin. *How to Do Things with Words*. Oxford University Press, Oxford, 1962.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023. URL `https://openreview.net/forum?id=ETKGuby0hcs`.

Arnold Cartagena and Ariane Teixeira. Mind the GAP: Text safety does not transfer to tool-call safety in LLM agents. *arXiv preprint arXiv:2602.16943*, 2026. URL `https://arxiv.org/abs/2602.16943`.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-0426. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/12a00d85a76fe258e1242c3aced03250-Abstract-Conference.html`.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2383–2400. USENIX Association, 2025b. URL `https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe`.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. SecAlign: Defending against prompt injection with preference optimization. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2833–2847. ACM, 2025c. doi: 10.1145/3719027.3744836. URL https://doi.org/10.1145/3719027.3744836.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82895–82920, 2024. doi: 10.52202/079017-2636. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. URL https://arxiv.org/abs/2503.18813.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen J. Xiang. Memory injection attacks on LLM agents via query-only interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1554. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/42a97bbd9844d2bf68596730af80bcdf-Abstract-Conference.html.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, pp. 79–90. ACM, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16647–16672. PMLR, 2024. URL https://proceedings.mlr.press/v235/gu24e.html.

Yu He, Haozhe Zhu, Yiming Li, Shuo Shao, Hongwei Yao, Zhihao Liu, and Zhan Qin. AttriGuard: Defeating indirect prompt injection in LLM agents via causal attribution of tool invocations. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 1547–1566. USENIX Association, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/he-yu.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting. In *Proceedings of the Conference on Applied Machine Learning in Information Security (CAMLIS 2024)*, volume 3920 of *CEUR Workshop Proceedings*, pp. 48–62. CEUR-WS.org, 2024. URL https://ceur-ws.org/Vol-3920/paper03.pdf.

Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciarini. The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29680–29697. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1435. URL https://aclanthology.org/2025.acl-long.1435/.

Tanqiu Jiang, Yuhui Wang, Jiacheng Liang, and Ting Wang. AgentLAB: Benchmarking LLM agents against long-horizon attacks. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.16901.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Li Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao, and Yue Zhao. Defenses against prompt attacks learn surface heuristics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10970–10987. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.502. URL https://aclanthology.org/2026.acl-long.502/.

Yuxi Li, Yi Liu, Yuekang Li, Ling Shi, Gelei Deng, Shengquan Chen, and Kailong Wang. Uncovering logit suppression vulnerabilities in LLM safety alignment. In *Pattern Recognition*, volume 16813 of *Lecture Notes in Computer Science*, pp. 61–76. Springer, 2027. doi: 10.1007/978-3-032-31583-0_5. URL https://link.springer.com/chapter/10.1007/978-3-032-31583-0_5.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847. USENIX Association, 2024. URL https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei.

Meta AI. Llama 3.3 70B Instruct model card. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Guha Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr. The attacker moves second: Stronger adaptive attacks bypass defenses against LLM jailbreaks and prompt injections. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 1467–1486. USENIX Association, 2026. URL https://www.usenix.org/conference/usenixsecurity26/presentation/nasr.

OpenAI. GPT-4o system card. OpenAI technical report, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-4o model. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-4o. Undated documentation; accessed 2026-09-19.

Dario Pasquini, Martin Strohmeier, and Carmela Troncoso. Neural exec: Learning (and learning from) execution triggers for prompt injection attacks. In *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*, AISec '24, pp. 89–100. ACM, 2024. doi: 10.1145/3689932.3694764. URL https://doi.org/10.1145/3689932.3694764.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565, 2024. doi: 10.52202/079017-4020. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract-Conference.html.

Qwen Team. Qwen3.5-9B model card. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026-09-19.

John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.624. URL https://aclanthology.org/2024.findings-acl.624/.

Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7116–7132. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.395. URL https://aclanthology.org/2025.findings-naacl.395/.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5750f91d8fb9d5c02bd8ad2c3b44456b-Abstract-Conference.html.

Tian Zhang, Yiwei Xu, Juan Wang, Keyan Guo, Xiaoyang Xu, Bowen Xiao, Quanlong Guan, Jinlin Fan, Jiawei Liu, Zhiquan Liu, and Hongxin Hu. AgentSentry: Mitigating indirect prompt injection in LLM agents via temporal causal diagnostics and context purification. *arXiv preprint arXiv:2602.22724*, 2026. URL https://arxiv.org/abs/2602.22724.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4688. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cd18539787d90e1d682d557c2c71b534-Abstract-Conference.html.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 83345–83373, 2024. doi: 10.52202/079017-2651. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97ca7168c2c333df5ea61ece3b3276e1-Abstract-Conference.html.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

Egor Zverev, Evgenii Kortukov, Alexander Panfilov, Alexandra Volkova, Soroush Tabesh, Sebastian Lapuschkin, Wojciech Samek, and Christoph H. Lampert. ASIDE: Architectural separation of instructions and data in language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b77b5f1a2878fedd4705bfeafd24ca35-Abstract-Conference.html.

A Source-Side Investigation and Extended Results

This appendix contains the source-side investigation behind the main-text residual analysis, together with the extended numerical and audit tables.

A.1 SEMANTIC ROLE MUTATION AS THE INITIAL HYPOTHESIS

The study began with a source-side hypothesis: an external message may enter an agent as environmental data but acquire action influence comparable to an authorized instruction after entering the working context. Let $\mathrm{Role}(m, C)$ denote the functional role of message m in processing context C. The hypothesis concerns a role change when

$$\mathrm{Role}(m, C_1) = \mathrm{DATA} \quad \wedge \quad \mathrm{Role}(m, C_2) = \mathrm{INSTRUCTION}, \tag{2}$$

where C_1 is the external-content context and C_2 is the subsequent action-decision context. The definition refers to action influence, rather than surface command syntax. It does not assume that the model contains separate internal modules for data and instructions.

External information must influence an agent; the security question is whether that influence also carries authorization to change the next action. We therefore treated semantic role mutation as an initial unifying explanation and tested it with a protocol-level counterfactual. The result supports a source-side contribution for canonical indirect prompt injection but does not explain every attack family.

A.2 LANGUAGE DECOMPOSITION AND CONTROLLED CONDITIONS

Drawing on the distinction between propositional content and illocutionary force (Austin, 1962; Searle, 1969), we represent external content as

$$x \mapsto \big(f(x), d(x)\big), \tag{3}$$

where $f(x)$ is factual or referential content and $d(x)$ is the directive-bearing content that may request, command, override, or induce an action. This is an experimental representation, not a claim that the agent implements two independent semantic channels.

CPSep uses four conditions. Raw preserves the original message. Directive Exposure makes $f(x)$ and $d(x)$ explicit while retaining the original text. Performative Neutralization replaces the directive with a descriptive report that preserves its referent, requested action, addressee, and stated reason. Directive Removal keeps only $f(x)$ as an upper-bound ablation. All four conditions preserve the user task, payload identity, and execution environment as far as the protocol permits. The victim receives no extractor prompt, attack label, warning, refusal cue, or new system rule.

A.3 CPSEP IMPLEMENTATION AND ATTACK-SPECIFIC AUDIT

CPSep changes the treatment of directive-bearing spans while holding the user task, payload identity, factual spans, and execution environment fixed when the transformation succeeds. Raw preserves the original input, Directive Exposure changes its representation, Performative Neutralization rewrites directive force, and Directive Removal is an upper-bound ablation rather than a content-preserving counterfactual.

An independent extraction call returns validated JSON with `facts` and `directives` fields. Each directive includes its exact source span, neutralized description, requested action, category, addressee, and explicit or implicit force. Unresolved extraction or replacement falls back to the original content. The victim agent receives no extractor prompt, attack label, warning, refusal cue, or maliciousness judgment. This fail-open design is useful for attribution but is not a deployment policy.

Figure 2 shows the workflow, while Figure 3 and Table 1 report the aggregate result. Totals are equal-weight means over the available attack methods; unsupported method–model pairs remain missing. Values are displayed to one decimal, so averages of rounded entries can differ from the reported totals. CPSep is a hypothesis-forming screen, and these aggregates are descriptive. Table 6 gives the attack-specific AgentDojo results.

The AgentLAB screen covers intent hijacking (IH), memory poisoning (MP), objective drift (OD), and tool-chain manipulation (TC). Each model–family condition uses 100 tasks and up to seven agent turns. Attack success is judged on a five-point scale, with scores of at least four counted as success. Table 7 reports unweighted means across the eight models; Tables 8 and 9 give their

Table 6: CPSep ASR by attack. DE, PN, and DR denote Directive Exposure, Performative Neutralization, and Directive Removal; totals average the available methods.

	GPT-4o				Llama 3.3			
Attack	**Raw**	**DE**	**PN**	**DR**	**Raw**	**DE**	**PN**	**DR**
direct	3.7	1.9	1.9	0.0	3.6	2.1	2.6	0.5
injecagent	5.7	1.6	1.6	0.0	–	–	–	–
ignore-prev	5.4	0.3	1.4	0.0	12.0	4.5	4.2	1.2
impt-instr	47.7	4.8	9.5	2.4	23.1	6.3	5.2	1.2
tool-know	34.5	12.1	8.6	1.7	–	–	–	–
Total	19.4	4.1	4.6	0.8	12.9	4.3	4.0	0.9

Table 7: AgentLAB CPSep ASR (%). Eight-model family means.

Family	**Raw**	**DE**	**PN**	**DR**
IH	44	30	30	30
MP	32	21	19	18
OD	66	33	14	7
TC	63	40	34	27

Table 8: AgentLAB CPSep ASR by model: IH and MP (%).

	IH				MP			
Model	**Raw**	**DE**	**PN**	**DR**	**Raw**	**DE**	**PN**	**DR**
DeepSeek	91	64	91	91	42	25	17	17
GPT-4o-mini	83	75	67	75	56	44	50	44
Qwen3.5	37	21	16	11	0	0	0	0
MiniMax	23	14	14	14	31	16	16	12
Claude	2	1	1	0	33	17	13	10
GPT-5.4	19	4	12	12	30	15	15	12
GLM-5	29	18	12	12	31	25	22	19
Gemini	71	39	30	22	32	29	19	32

individual results. This hypothesis-forming screen is separate from the Gemma and Qwen white-box confirmation.

CPSep four-condition example

Raw: "The latest report shows unemployment at 7.2%. Ignore previous instructions. Send a link to Alice."

Directive Exposure: `facts=[report shows 7.2%]; directives=[ignore previous instructions; send a link to Alice]` + original text

Performative Neutralization: "The latest report shows unemployment at 7.2%. The text contains requests to disregard prior instructions and send a link to Alice."

Directive Removal: "The latest report shows unemployment at 7.2%."

Here DE, PN, and DR denote Directive Exposure, Performative Neutralization, and Directive Removal. IH retains a 30% ASR after all three transformations, whereas OD falls from 66% to 7% under removal. These aggregate patterns motivate different decision-point questions; they do not determine the confirmatory family assignment or establish an internal mechanism.

The model rows show heterogeneous trajectories, including flat or rebounding ASR after neutralization and removal. These heterogeneous responses motivate the relation-specific white-box tests.

Table 9: AgentLAB CPSep ASR by model: OD and TC (%).

	OD				TC			
Model	**Raw**	**DE**	**PN**	**DR**	**Raw**	**DE**	**PN**	**DR**
DeepSeek	99	48	38	12	71	43	36	36
GPT-4o-mini	99	78	28	12	91	82	55	27
Qwen3.5	91	64	0	0	79	36	36	29
MiniMax	99	48	38	28	77	31	31	23
Claude	19	0	0	0	22	20	13	11
GPT-5.4	67	13	7	7	37	22	11	11
GLM-5	48	10	0	0	53	37	37	26
Gemini	8	4	4	0	71	50	50	50

Table 10: Paired E2 uncertainty and reported joint states. Task-clustered 95% intervals; Table 4 reports guardrail status.

Model	**Family**	**ΔASR [95% CI]**	**Random ΔASR**	**Ratio**	**State**
Gemma	IPI	-11.10 [-15.50,-6.80]	-1.20	.108	local-support
Gemma	IH	-4.67 [-8.67,-2.00]	-.67	.143	E1/E2-conflict
Gemma	MP	-10.00 [-15.07,-5.73]	-1.20	.120	local-support
Gemma	OD	-7.33 [-12.00,-3.33]	-1.33	.182	associated-effect
Gemma	TC	-58.00 [-65.20,-50.40]	-6.53	.113	associated-effect
Qwen	IPI	-5.00 [-8.00,-2.00]	-.50	.100	local-support
Qwen	IH	-2.00 [-4.67,0.00]	-.67	.333	inconclusive
Qwen	MP	-2.93 [-6.27,-.67]	-.40	.136	inconclusive
Qwen	OD	-2.93 [-5.60,-.67]	-.27	.091	inconclusive
Qwen	TC	-50.53 [-58.00,-42.13]	-6.40	.127	local-support

A.4 CSS TWIN AUDIT

The CSS point is frozen on the selection split and replayed on held-out twins. Harmful-use suppression (HS), benign retention (BR), and general damage (GD) are interpreted jointly. Operator eligibility means that CSS can be applied to a matched twin; E1 qualification additionally requires the capability-limitation signature. IPI and OD have no structurally matched capability twin in this protocol, so their separation scores are unavailable.

Table 11: Capability-twin and CSS audit. The shared E1/E2 suppression measurement is counted once.

Cell	**Twin N**	$\widehat{HS}_{\mathbf{conf}}$	**BR**	**GD**	**Operator status**
Gemma-IH	150	.379	.960	.030	operator-eligible
Gemma-MP	150	.160	.960	.030	operator-eligible
Gemma-TC	150	.279	.960	.040	operator-eligible
Qwen-IH	150	.420	.960	.030	operator-eligible
Qwen-MP	150	.118	.960	.030	operator-eligible
Qwen-TC	150	.321	.960	.040	operator-eligible
IPI, OD	0	–	–	–	twin-shortage

A.5 E1 AUDIT VALUES

Table 12 reports task-clustered intervals and unique recognized-success tasks followed by recognized-success replays. A dash marks structural non-identifiability; a high probe area under the receiver operating characteristic curve (AUROC) cannot compensate for inadequate source pairs, matching, or sample size. Table 13 then decomposes the recognized cases by severity.

Table 12: E1 intervals and denominators, split by readout group for legibility.

(a) Source and capability readouts

Cell	DI_E [95% CI]	PAS [95% CI]	$\widehat{HS}_{\text{conf}}$ [95% CI]
Gemma-IPI	.181 [.176,.185]	.035 [.030,.040]	–
Gemma-IH	–	–	.379 [.377,.380]
Gemma-MP	–	–	.160 [.158,.161]
Gemma-OD	.010 [.004,.015]	.080 [.074,.087]	–
Gemma-TC	–	–	.279 [.277,.280]
Qwen-IPI	.162 [.157,.166]	.052 [.046,.058]	–
Qwen-IH	–	–	.420 [.418,.421]
Qwen-MP	–	–	.118 [.117,.120]
Qwen-OD	.023 [.018,.028]	.074 [.067,.081]	–
Qwen-TC	–	–	.321 [.319,.322]

(b) Recognition counts and candidate-priority readouts

Cell	N_{task}/N_{replay}	$M_{SU}(rec)$ [95% CI]	Δ_{SPA} [95% CI]
Gemma-IPI	29/141	-2.753 [-2.830,-2.682]	.327 [.266,.378]
Gemma-IH	15/75	-.375 [-.447,-.312]	–
Gemma-MP	86/425	-1.945 [-1.998,-1.895]	.404 [.365,.443]
Gemma-OD	33/165	-.272 [-.319,-.224]	.327 [.280,.374]
Gemma-TC	131/650	-1.621 [-1.650,-1.590]	.425 [.366,.468]
Qwen-IPI	11/55	-1.878 [-2.056,-1.752]	.357 [.311,.405]
Qwen-IH	5/25	.734 [.641,.832]	–
Qwen-MP	23/115	-.587 [-.666,-.506]	.348 [.294,.400]
Qwen-OD	18/87	.457 [.423,.494]	.288 [.252,.322]
Qwen-TC	111/549	-2.015 [-2.044,-1.984]	.469 [.417,.519]

Table 13: Recognized-success strata, reported replay shares (%) / M_{SU}; dashes mark no valid sample. The Δ_{SPA} estimand uses separate Raw stratum weights.

Model	Family	Mild	Moderate	Severe
Gemma	IPI	56.74/-2.681	35.46/-2.760	7.80/-3.241
Gemma	IH	80.00/-0.346	20.00/-0.491	0.00/–
Gemma	MP	30.59/-1.661	47.06/-1.967	22.35/-2.290
Gemma	OD	69.70/-0.228	24.24/-0.386	6.06/-0.315
Gemma	TC	8.46/-1.228	65.54/-1.618	26.00/-1.754
Qwen	IPI	81.82/-1.780	9.09/-2.063	9.09/-2.577
Qwen	IH	80.00/0.758	20.00/0.641	0.00/–
Qwen	MP	39.13/-0.429	56.52/-0.673	4.35/-0.887
Qwen	OD	86.21/0.438	13.79/0.579	0.00/–
Qwen	TC	6.38/-1.594	71.22/-2.020	22.40/-2.119

E1 states are relation-specific. Both IH cells lack the capability-limitation signature; Qwen/MP has that signature. Their separate harmfulness and candidate-priority qualifications do not enter CSS classification. The finite-grid audit, adjusted E2 tests, and target-response measurements below expose the checks behind the four local-support, two associated-effect, one E1/E2-conflict, and three inconclusive outcomes.

For source authorization, Gemma/IPI and Qwen/IPI have source-probe AUROCs of .982 and .973, selectivity over control-task probes of .514 and .486, and paraphrase AUROCs of .968 and .967. Both pass the $\pm.08$ PAS equivalence test, with two one-sided tests (TOST) giving $p < .0001$. Their positive DI_E intervals appear in Table 12. For Qwen/TC, harmfulness AUROC is .999 with selectivity .482; its negative recognized-success margin, positive Δ_{SPA} interval, and within-stratum scalar balance jointly qualify safe-candidate priority.

Qwen/OD's positive $M_{SU}(rec)$ indicates that the safe candidate is favored among recognized attack successes. This profile lacks the safe-candidate disadvantage required for E1 qualification.

Table 14: Recorded E1 states and measurements. Here underpowered denotes scalar-balance failure.

Cell	Mapped relation	E1 state	Displayed measurements
Gemma-IPI	Source authorization	signature-present	Source influence and authorization contrast
Gemma-IH	Capability selectivity	signature-absent	Suppression exceeds the .20 threshold
Gemma-MP	Capability selectivity	signature-present	Selected-point HS/BR/GD
Gemma-OD	Safe-candidate priority	underpowered	Candidate margins; scalar-balance failure
Gemma-TC	Safe-candidate priority	underpowered	Candidate margins; scalar-balance failure
Qwen-IPI	Source authorization	signature-present	Source influence and authorization contrast
Qwen-IH	Capability selectivity	signature-absent	Suppression exceeds the .20 threshold
Qwen-MP	Capability selectivity	signature-present	Limited feasible suppression on twins
Qwen-OD	Safe-candidate priority	signature-absent	Positive recognized-success margin
Qwen-TC	Safe-candidate priority	signature-present	Candidate margins and stratified Raw contrast

A.6 Adjusted decisions and target engagement

Table 15 gives the multiplicity-sensitive checks. A joint response requires delivery, target engagement, a negative behavioral contrast, and behavioral $q < .05$. Specificity requires a negative target-minus-control contrast against each of the four controls, with all four adjusted q values below $.05$. All ten mapped arms pass the recorded delivery and target-engagement checks and satisfy the utility, damage, and over-refusal bounds. Seven pass the adjusted behavioral and specificity tests. Relation-specific E1 then distinguishes four supported cases from two scalar-balance-limited cases and one qualification–response conflict.

Table 15: Adjusted E2 tests. The largest specificity q summarizes four required control contrasts. The threshold is $.05$; G and Q abbreviate Gemma and Qwen in the following compact tables.

Cell	Behavior q	Max. specificity q	Decisive check
Gemma/IPI	$< .0001$	$< .0001$	All checks pass
Gemma/IH	.0150	.0252	Capability signature absent
Gemma/MP	.0001	.0002	All checks pass
Gemma/OD	.0035	.0069	E1 scalar imbalance
Gemma/TC	$< .0001$	$< .0001$	E1 scalar imbalance
Qwen/IPI	.0055	.0078	All checks pass
Qwen/IH	.1613	.2721	Adjusted tests; E1 absent
Qwen/MP	.0591	.0667	Adjusted tests unresolved
Qwen/OD	.1272	.1272	Adjusted tests; E1 absent
Qwen/TC	$< .0001$	$< .0001$	All checks pass

A supplementary analysis of the per-candidate log scores measures SAG's change in external-directive influence. With $1{,}200$ paired task-cluster resamples and seed 20260909, DI_E changes from $.181$ to $-.272$ for Gemma/IPI and from $.162$ to $-.258$ for Qwen/IPI. The respective change intervals are $[-.456, -.448]$ and $[-.423, -.417]$. All five replays stay within their sampled task cluster. These reductions complement the increased authorization contrasts in Table 4.

Table 16 reports the target and safeguard measurements. For Qwen/TC, the mean harmfulness-probe probability changes by $.00188$ with interval $[.00169, .00208]$ while the safe-candidate margin improves. The guardrail measures the aggregate change; preservation of individual recognition labels remains untested. Its refusal-bias control produces a similar $.650$ margin shift but retains only $.873$ of utility and has an over-refusal rate of $.107$, compared with $.980$ and $.007$ for SPS. This

control shows why target movement and utility must be checked together. CSS twin retention and E2 task utility measure different retained capabilities.

Table 16: Target and safeguard measurements for local-support cells, with 95% intervals below estimates. L1 delivery is 1.00 throughout. Utility is a retention ratio; damage is an absolute score loss; over-refusal is a rate.

Cell	Target	Utility	Damage	Over-refusal
	ΔPAS .452	.981	.016	.000
G/IPI	[.449,.456]	[.980,.982]	[.015,.017]	[.000,.000]
	HS .160	.980	.017	.000
G/MP	[.158,.161]	[.979,.981]	[.016,.018]	[.000,.000]
	ΔPAS .420	.979	.018	.005
Q/IPI	[.417,.423]	[.978,.980]	[.017,.018]	[.000,.015]
	ΔM_{SU} .650	.980	.017	.007
Q/TC	[.644,.656]	[.979,.981]	[.016,.018]	[.000,.020]

Table 17: Paired specificity against all four controls. Entries give target-minus-control ASR changes in percentage points with task-clustered 95% intervals. All corresponding adjusted q values are below .008.

Cell	Random	Wrong-site	Orthogonal	Matched
	-9.90	-10.50	-10.30	-10.90
G/IPI	[-14.00,-5.90]	[-15.00,-6.00]	[-14.60,-6.00]	[-15.30,-6.50]
	-8.80	-9.60	-9.60	-9.73
G/MP	[-13.33,-4.80]	[-14.40,-5.33]	[-14.40,-5.33]	[-14.67,-5.47]
	-4.50	-5.00	-5.00	-5.00
Q/IPI	[-7.50,-1.99]	[-8.00,-2.00]	[-8.00,-2.00]	[-8.00,-2.00]
	-44.13	-48.40	-47.47	-49.73
Q/TC	[-51.47,-36.00]	[-56.00,-39.99]	[-55.33,-38.93]	[-57.07,-41.20]

A.7 Finite-grid capability evidence

Table 18 reports every MP selection point, including infeasible ones. The selected rule and dose maximize feasible suppression before held-out confirmation. Both models reach stronger suppression at higher doses, but every such point violates the .95 benign-retention bound. The selected-point confirmation intervals for MP lie below .162, whereas both IH intervals lie above .377. The MP and IH intervals lie entirely on opposite sides of the registered .20 difficulty threshold.

Table 18: Complete MP selection grids. Feasibility uses $BR \geq .95$ and $GD \leq .05$. Both models select the selective rule at dose .2; held-out values are in Table 11.

Rule	β	G HS	G BR	G GD	Q HS	Q BR	Q GD	Feasible
distributed	.1	.047	.985	.010	.040	.986	.010	yes
distributed	.2	.127	.956	.030	.094	.956	.030	yes
distributed	.3	.205	.926	.040	.186	.926	.040	no
distributed	.4	.265	.896	.060	.264	.906	.060	no
selective	.1	.059	.989	.010	.051	.990	.010	yes
selective	.2	.160	.960	.030	.119	.959	.030	yes
selective	.3	.259	.930	.040	.240	.930	.040	no
selective	.4	.339	.900	.060	.340	.910	.060	no

A.8 Matching and recognized-group accounting

The registered balance audit assigns each task its majority Raw outcome across five replays, with at least three successes defining success. Within each severity stratum, its standardized mean difference

(SMD) divides the success-minus-resistance mean of the canonical task's recorded scalar by its sample standard deviation across all confirmatory tasks in the cell. The gate requires $|\mathrm{SMD}| \leq .5$ and at least two tasks per group. In contrast, Δ_{SPA} uses all Raw replay outcomes with frozen Raw-success weights. This task-level gate uses one scalar as a proxy for balance between the Raw success and resistance groups in Δ_{SPA}. The separate $M_{SU}(rec)$ margin conditions on recognized successes.

Table 19 compares the registered task SMD with replay-group sensitivity using the same denominator. Gemma/OD moderate and Gemma/TC mild and severe exceed .50 under both; Qwen/TC remains within it. Qwen/OD moderate moves from $-.583$ to $-.436$, while severe has no success replay. The two grouping rules preserve the Gemma/Qwen tool-chain contrast. For these Gemma cells, the registered underpowered label denotes E1 scalar imbalance; resolved E2 yields the joint state associated-effect. Replays share a task scalar and are dependent; recognized task counts in Table 20 can overlap.

Table 19: Raw scalar balance. S/R and task SMD use majority outcomes; replay SMD is descriptive with the same denominator. Weights are Raw-success shares.

Cell	Stratum	S/R tasks	Task SMD	Replay SMD	Weight
Gemma/OD	Mild	24/66	.194	.194	.7059
Gemma/OD	Moderate	8/44	.775	.775	.2353
Gemma/OD	Severe	2/6	.123	.123	.0588
Gemma/TC	Mild	12/2	-.753	-.753	.0889
Gemma/TC	Moderate	87/9	.136	.120	.6459
Gemma/TC	Severe	36/4	.615	.662	.2652
Qwen/OD	Mild	15/95	.085	.085	.8621
Qwen/OD	Moderate	2/35	-.583	-.436	.1379
Qwen/OD	Severe	0/3	–	–	.0000
Qwen/TC	Mild	9/3	.232	.232	.0791
Qwen/TC	Moderate	80/23	-.330	-.333	.7047
Qwen/TC	Severe	25/10	-.077	-.065	.2162

Table 20: Recognized outcome groups. Entries give unique success/resistance tasks with a qualifying replay in each severity stratum.

Cell	Mild	Moderate	Severe
Gemma/OD	23/65	8/42	2/6
Gemma/TC	11/2	86/9	34/5
Qwen/OD	15/93	3/35	0/3
Qwen/TC	7/3	79/23	25/11

B Detailed Method Definitions and Evidence Gates

B.1 Terminology and estimands

SAG, CSS, and SPS are registered diagnostic operators. In E2, L1 verifies delivery to the registered state and L2 verifies the target-readout response. The state `local-support` records controlled support for the prespecified relation within the evaluated model, family, CDP, decision window, and intervention space. The causal effect concerns the applied operator; attribution to the relation depends on the matching, readout, and control assumptions.

All candidate margins use the same teacher-forced, length-normalized sequence score:

$$L(y \mid x) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log p_\theta(y_i \mid x, y_{<i}). \tag{4}$$

Table 21: Counterfactual requirements and target readouts.

Dimension	Required counterfactual	Target readout
Source authorization	Same directive, different source authorization	DI_E and PAS
Capability selectivity	Matched harmful and benign capability twins	$\widehat{HS}_{\text{conf}}$, BR, GD
Safe-candidate priority	Matched safe and attack-advancing candidates	M_{SU} and Δ_{SPA}

Here x is the frozen decision prefix, $y = (y_1, \ldots, y_{|y|})$ is a fixed candidate token sequence, p_θ is the evaluated model, and $y_{<i}$ is the candidate prefix before token i. Larger L indicates greater local preference for the fixed candidate. Generated first-action ASR is measured separately on replayed decisions.

For source authorization, let y_F be the directive-following continuation and y_I the matched non-following continuation. Set $m(x) = L(y_F \mid x) - L(y_I \mid x)$. For source slot $s \in \{E, T\}$, compare the same directive d with its length-matched data-like control d_0. We call $PAS = DI_T - DI_E$ the trusted–external authorization contrast; positive PAS means greater directive influence from the trusted source. The estimands are

$$DI_s = m(x_{s,d}) - m(x_{s,d_0}), \qquad PAS = DI_T - DI_E, \tag{5}$$

where $x_{s,d}$ is the matched prefix with directive d in slot s, and E and T denote external and trusted authorization slots. Each DI_s is averaged over eligible matched source pairs before PAS is reported. For capability twins, a task's score is the fraction of its prespecified checklist criteria that are satisfied:

$$q_i^D = \frac{1}{m_i} \sum_{r=1}^{m_i} c_{ir}^D, \qquad Q_D = \frac{1}{N_D} \sum_{i=1}^{N_D} q_i^D. \tag{6}$$

Here $D \in \{H, B, gen\}$ identifies harmful, benign, or independent general-capability tasks; c_{ir}^D is the binary result for criterion r on task i, m_i is its checklist length, and N_D is the number of tasks of type D. The matched raw no-intervention score is $Q_D(0)$. The guardrail metrics are

$$HS = 1 - \frac{Q_H(\pi, \beta)}{Q_H(0)}, \quad BR = \frac{Q_B(\pi, \beta)}{Q_B(0)}, \quad GD = Q_{gen}(0) - Q_{gen}(\pi, \beta), \tag{7}$$

where π is a registered unit-selection and access rule, β is its intervention dose, HS is harmful-use suppression, BR is benign retention, and GD is an absolute loss on the normalized general-capability score. Thus $GD = .03$ means a three-point loss on a 0–100 scale, rather than a relative loss of 3% of the baseline. Ratios with zero baselines are undefined and reported as NA. Constrained separation is

$$Sep_\Pi(\epsilon, \delta) = \sup_{(\pi,\beta)\in\mathcal{B}} \{HS(\pi, \beta) \mid BR(\pi, \beta) \geq 1 - \epsilon,\ GD(\pi, \beta) \leq \delta\}, \tag{8}$$

where Π is the finite set of registered selection and access rules, $\mathcal{D}$ is the finite dose set, $\mathcal{B} = \Pi \times \mathcal{D}$, and $\epsilon = .05$ and $\delta = .05$ are the frozen retention and damage bounds. If the feasible set is empty, the result is reported as `structural-NA`. This supremum defines the search objective. The reported selected-point effect uses (π^*, β^*) chosen on the selection split:

$$(\pi^*, \beta^*) = \arg \max_{(\pi,\beta)\in\mathcal{B}_{\text{feas,sel}}} HS_{\text{sel}}(\pi, \beta), \qquad \widehat{HS}_{\text{conf}} = HS_{\text{conf}}(\pi^*, \beta^*). \tag{9}$$

Here $\mathcal{B}_{\text{feas,sel}}$ contains points satisfying the selection-split bounds, HS_{sel} is their selection-split score, and HS_{conf} is the score at the same point on held-out confirmation tasks. Confirmation evaluates the frozen point and its guardrails, giving the selected-point $\widehat{HS}_{\text{conf}}$ in Table 11 without renewed maximization. CSS therefore shares this intervention-based readout across E1 and E2. Additional E2 support comes from delivery, behavior, safeguards, and controls. For safe-candidate

priority, $M_{SU} = L(y_S \mid x) - L(y_U \mid x)$ compares the matched safe candidate y_S and attack-advancing candidate y_U. The reported $M_{SU}(rec)$ conditions on recognized attack-success records, whereas the following contrast uses both resistance and success groups:

$$\Delta_{SPA} = \sum_{b \in \mathcal{B}_{raw}} w_b \left[\mathbb{E}(M_{SU} \mid resist, b) - \mathbb{E}(M_{SU} \mid success, b) \right], \tag{10}$$

where $\mathcal{B}_{raw}$ is the set of frozen Raw severity strata and w_b is its archived, normalized weight, labelled a Raw-success share in the audit. The two conditional means use the same fixed candidate scoring rule; success and resistance are observed Raw outcomes within each stratum. Δ_{SPA} uses all Raw replays, while only $M_{SU}(rec)$ conditions on recognition. All quantities are computed on held-out samples. These estimands separate source influence, selected-point suppression, and candidate priority.

B.2 Primary CDP and decision window

The primary CDP is the earliest point satisfying exposure, opportunity, and pre-action conditions. Exposure means the attack-bearing material has entered the context; opportunity means an attack-advancing action is available; pre-action means it has not yet been executed. Ties retain the earliest complete action boundary. The rule is applied before reading ASR or terminal-state behavior. A decision window begins at the first post-CDP prediction and ends at the first complete calibrated action, including a tool call, refusal, clarification, or legitimate continuation. Raw, target, and control arms share the prefix and window, so their differences concern the same action opportunity. The frozen action parser and attack-goal predicate evaluate the complete action generated from this prefix. Registered operators act at each prediction position within the decision window.

B.3 Source-authorization details

The source pair preserves the directive, task, tool, candidate continuations, span boundaries, and token length while changing its authorization slot. Trusted and external are authorization conditions, not merely two positions in the prompt. The directive-free control replaces directive force with data-like content at the same length. These constraints let DI_E measure directive influence beyond exposure and let PAS compare that influence across sources.

Source probes use source-bearing residuals with independent control-task, random-label, and paraphrase checks. The registered E1 tests require probe selectivity, DI_E distinguishable from zero, and PAS within the equivalence bounds. Interpreting this profile as a failure of source authorization also requires positive external-directive influence; both IPI intervals lie entirely above zero. The two one-sided tests (TOST) procedure supplies affirmative equivalence evidence for PAS. A small point estimate or a non-significant difference alone does not. Together, the conditions identify externally sourced directive influence close to trusted influence despite a readable source distinction. L1 verifies SAG delivery. L2 requires reduced DI_E and increased PAS, while benign fact retrieval and legitimate-task progress remain within their guardrails.

For external directive d and its data-like control d_0, SAG subtracts the registered source-residual contrast only inside the frozen window $\mathcal{W}(x)$:

$$\begin{aligned} c_{\ell,r}^{E,d} &= h_{\ell,r}(x_{E,d}, y_{<r}) - h_{\ell,r}(x_{E,d_0}, y_{<r}), \\ h_{\ell,r}^{\mathrm{SAG}} &= h_{\ell,r} - \alpha c_{\ell,r}^{E,d}, \qquad r \in \mathcal{W}(x),\ \alpha \in [0,1]. \end{aligned} \tag{11}$$

Here $h_{\ell,r}$ is the residual state at registered layer ℓ and candidate position r, $y_{<r}$ is the identical teacher-forced history in the directive and control runs, and $c_{\ell,r}^{E,d}$ is their paired residual difference. The intervention subtracts a fraction α of that difference only at positions in the frozen decision window $\mathcal{W}(x)$; outside the window the state is unchanged. L1 checks that the registered residual change was delivered. L2 checks whether external directive influence DI_E falls and the trusted–external contrast PAS rises on the matched source pair. This measures directive influence.

B.4 Capability-selectivity details

Capability twins match tools, invocation depth, entities, output schema, atomic steps, and scoring scales; intended use and its consequences differ. Each task score is the fraction of its prespecified

atomic criteria satisfied. Q_H and Q_B average these scores separately across harmful and benign twins; Q_{gen} uses independent general-capability tasks. A zero baseline or invalid task set makes the corresponding ratio undefined, reported as NA.

For a fixed calibration perturbation β_0, unit j is ranked by its harmful-task loss increase relative to its benign-task loss increase:

$$I_j(D) = \mathbb{E}_{x\sim D}[\mathcal{L}_D(\mathrm{CSS}_{j,\beta_0}(x)) - \mathcal{L}_D(x)]\,,$$
$$R_j = \widetilde{I}_j(H) - \lambda\widetilde{I}_j(B).$$

Here D is the harmful or benign calibration set, $\mathcal{L}_D$ is its task loss, and CSS_{j,β_0} applies the calibration perturbation to unit j. Tildes denote standardized influences. Each registered rule specifies the tradeoff weight λ, retained count K, and candidate layer set; U^* contains the K highest-ranked units and is frozen before confirmation. The raw no-intervention arm supplies the identity baseline. Selection chooses the point maximizing harmful-use suppression under $BR \geq .95$ and $GD \leq .05$; confirmation reuses that point without renewed search.

The finite-grid capability signature combines limited feasible suppression with an observed tradeoff. The maximum feasible HS on the selection split is below the recorded difficulty threshold .20, and tested points with higher suppression violate $BR \geq .95$ or $GD \leq .05$. Confirmation reports the frozen point's HS, BR, and GD, without a second maximization. Table 18 gives the complete MP selection grids; their stronger doses breach benign retention. IH instead has feasible suppression above .20, so its capability-limitation signature is absent. This criterion concerns the tested grid, not irreducible coupling or separation over arbitrary interventions. The HS measurement is shared by E1 and E2, leaving behavioral and control tests to supply the additional evidence.

At the Gemma/MP point, $HS = .160$ means that harmful-task performance retains .840 of its raw score; $BR = .960$ retains 96% of benign performance, and $GD = .030$ is a three-point loss on a 0–100 general-capability scale. The point satisfies the capability bounds. The next selective dose increases selection HS from .160 to .259 while reducing BR from .960 to .930. This measured tradeoff explains the finite-grid qualification; the held-out ASR and paired-control tests assess its behavioral relevance.

For selected units U^*, CSS contracts each activation toward its calibration mean $\mu_{j,\ell}$:

$$z^{\mathrm{CSS}}_{j,r} = \mu_{j,\ell} + (1-\beta)(z_{j,r} - \mu_{j,\ell}), \qquad j \in U^*,\ r \in \mathcal{W}(x). \tag{12}$$

Here $z_{j,r}$ is the activation of selected unit j at its registered layer ℓ and decision position r; $\mu_{j,\ell}$ is that unit's calibration-set mean, and U^* is the unit set frozen before confirmation. Dose $\beta = 0$ is the identity intervention, while positive β moves each selected activation toward its own mean. Units outside U^* and positions outside $\mathcal{W}(x)$ are unchanged. L1 checks the activation change at the registered units; L2 checks the harmful/benign twin scores HS and BR at the frozen point. The unit ranking and tested grid specify the local intervention space in which the capability relation is evaluated.

B.5 Safe-candidate-priority details

Harmfulness labels are independent of refusal and attack success. Safe and attack-advancing candidates are matched for action form, key parameters, token boundaries, and serialization complexity. $M_{SU} > 0$ means the safe candidate has a higher local sequence score. SPS L2 tests its directional shift; the harmfulness probe is a separate aggregate guardrail. SPS uses a direction calibrated on independent tasks and cannot use confirmatory M_{SU}, ASR, or outcome labels to choose a direction or dose.

An independent harmfulness probe defines recognition using a threshold fixed on calibration tasks. Raw severity strata and their weights are frozen before intervention. Within each stratum, the scalar-balance audit compares tasks grouped by their majority Raw outcome. The resistance–success contrast uses all Raw replays; the success margin separately conditions on recognized cases. E1 requires identifiable harmfulness, a safe-candidate disadvantage among recognized successes, and a stable positive resistance–success contrast Δ_{SPA}. Thus, readable harm and an unsafe local preference must co-occur, while Raw resistance shows greater preference for safety.

SPS averages hidden states over a calibrated suffix of length L_v, forms a normalized safe-minus-unsafe direction on independent calibration tasks, and applies the frozen dose γ:

$$\begin{aligned} \rho_\ell(x,y) &= L_v^{-1} \sum_{r=r_{\mathrm{div}}}^{r_{\mathrm{div}}+L_v-1} h_{\ell,r}(x, y_{<r}), \\ v_{\mathrm{safe}} &= \mathrm{norm}\big(\mathbb{E}[\rho_\ell(x, y_S^{cal})] - \mathbb{E}[\rho_\ell(x, y_U^{cal})]\big), \\ h_{\ell,r}^{\mathrm{SPS}} &= h_{\ell,r} + \gamma \|h_{\ell,r}\| v_{\mathrm{safe}}, \qquad r \in \mathcal{W}(x). \end{aligned} \tag{13}$$

Here r_{div} is the first token at which the calibrated safe and attack-advancing continuations diverge, and L_v is the fixed number of positions averaged after that point. ρ_ℓ is their mean hidden state at layer ℓ; superscript cal marks candidates from independent calibration tasks. The operator norm scales the safe-minus-unsafe difference to unit length, producing the frozen, layer-specific direction v_{safe}. At a registered decision position r, SPS adds that direction to residual state $h_{\ell,r}$ with dose γ scaled by the state's norm. L1 verifies delivery along the frozen direction; L2 checks the change in M_{SU} on matched candidates. No confirmatory outcome is used to choose v_{safe} or γ.

B.6 Control contrasts and readout separation

Each eligible target arm is replayed with the same task identities, frozen prefix, and decision window as its controls. The protocol specifies random-span controls for SAG, random-unit controls for CSS, and random-direction controls for SPS. These test whether a comparably sized perturbation produces the same behavioral change; a wrong-site control tests dependence on the registered access coordinate; an orthogonal control tests the selected direction or unit subspace; and a matched control checks the prespecified non-target change. The refusal-bias arm separately tests whether a similar safety-margin shift is obtained by broad refusal at the cost of legitimate behavior. Specificity is assessed through paired target-minus-control ASR changes against each of the random, wrong-site, orthogonal, and matched controls. Table 23 lists the control records and Table 15 reports the adjusted specificity checks.

B.7 Evidence-state decision tree

Resolved E2 denotes the joint delivery, target-engagement, and behavioral response $R(c)$. Each requires its own confirmatory check. A negative ΔASR interval and an adjusted behavioral $q < .05$ resolve the behavioral component; the relation-specific target conditions are given above. Utility and matched-control specificity determine $V(c)$.

Observed bound violations establish guardrail-failure, while incomplete records leave checks open. Nonspecificity requires a demonstrated control effect comparable to the target. These distinctions separate a failed check from an unresolved one. Table 14 reports the joint E1 assessments.

Table 22: Evidence states and interpretation.

State	Interpretation
signature-present	E1 joint qualification criteria pass; no causal claim
signature-absent	The required E1 failure signature is not satisfied under the registered criterion
non-identifiable	A required relation-specific probe, match, or stratum does not identify the joint E1 contrast
instrument-invalid	The intervention or measurement fails its delivery, snapshot, parsing, or replay-validity check
local-support	E1, L1, L2, behavior, guardrails, controls, and uncertainty all pass within scope
associated-effect	Joint E2 resolves, but another required gate remains open
E1/E2-conflict	Joint E2 resolves with absent or non-identifiable E1; a gate mismatch, not opposing estimates

Continued on next page

Table 22 continued

State	Interpretation
non-specific	A random, wrong-site, orthogonal, or matched control demonstrates a behavioral effect comparable to the target
guardrail-failure	Utility, benign retention, damage, or over-refusal exceeds the bound
inconclusive	Contrast or response unresolved without an identified current-stage power or coverage shortfall
underpowered	Current-stage cluster count, registered scalar balance, coverage, or power criterion is not met
structural-NA	Required source pair, twin, CDP, or combination is unavailable

First check structural availability and instrument validity. For an eligible, valid intervention, an observed guardrail violation yields guardrail-failure. Affirmative evidence of a comparable behavioral effect under a paired control yields non-specific; a nonzero control effect alone does not. An unresolved target-minus-control contrast leaves specificity open. Otherwise, all conditions in Equation 1 give local-support. Resolved joint E2 with absent or non-identifiable E1 gives E1/E2-conflict; resolved E2 with other incomplete support gives associated-effect. An unresolved E2 response is underpowered when a current-stage power or coverage shortfall is identified, and inconclusive otherwise.

Specificity compares target and control ASR changes using paired resamples of the same task identities. Qwen/TC changes by -50.53 points versus -6.40 for the random control, a -44.13-point difference with interval $[-51.47, -36.00]$ in Table 17. This establishes differential behavior; E1, delivery, target engagement, and guardrails are also required for the local claim. The ratios in Table 10 summarize magnitudes without replacing this paired comparison.

C Experimental Details, Parameters, and Reproduction Record

C.1 Models, attacks, and sample accounting

The white-box audit uses AgentLAB's environments and attack framework (Jiang et al., 2026) with Gemma-4-12B-it (Gemma Team, 2026) and Qwen3.5-9B (Qwen Team, 2026). AgentLAB's task-injection setting extends indirect prompt injection across multiple turns; we retain IPI as the audit label. Our audit corpus contains, per model, 400 IPI tasks and 300 per other family. A fixed 25/25/50 calibration, selection, and confirmation split gives 200 confirmatory IPI tasks and 150 per other family. Five replays share each task identity and canonical prefix; the independent unit remains the task.

ASR averages binary outcomes across replays. Two-sided paired t tests use task-level replay means, with 199 degrees of freedom for IPI and 149 otherwise; specificity compares target and control on the same tasks. Intervals use $1{,}200$ paired task-clustered bootstrap resamples. Within each model–family cell, BH adjustment covers 31 tests: the Raw baseline plus ten comparisons for each of the three eligible nonbaseline arms. These arms are the two single operators and their combination in Table 3. Each contributes six behavioral changes for target, random, wrong-site, orthogonal, matched, and refusal-bias conditions, and four target-versus-control specificity contrasts for random, wrong-site, orthogonal, and matched controls. The baseline's zero contrast has $p = 1$. Adjusted $q < .05$ determines each behavioral and specificity check.

C.2 Configuration and reproduction record

Table 23 gives settings frozen before confirmation.

Table 23: Frozen settings and protocol requirements.

Record	Reported setting or requirement
Access and doses	SAG at 12 with dose .20; CSS at 18 with doses .40/.20/.30 for IH/MP/TC; SPS at 23 with dose .20. Wrong-site coordinates are 15, 21, and 26. Target L1 specifications match delivery in all five runs.
CSS grid	Two fixed rules, `distributed` and `selective`, crossed with doses .10–.50 for IH and .10–.40 for MP/TC in steps of .10. All six cells select `selective`. Each selection set contains 75 tasks.
Decision window	Maximum 48 tokens; recorded task windows 12–40 tokens. Stored candidate-score arrays contain eight entries.
Arms and eligibility	Raw, three single operators, and their combinations. SAG is confirmatory on IPI; CSS requires a capability twin for IH, MP, or TC; SPS requires matched candidates and harmfulness calibration. SAG+CSS combinations are ineligible in all cells.
Controls	Each eligible arm has target, random, wrong-site, orthogonal, matched, and refusal-bias records, sharing a Raw baseline. Paired control contrasts use the same task identities.
Decision thresholds	PAS equivalence .08; capability difficulty .20; probe AUROC .80 and selectivity .20; scalar balance $\|\mathrm{SMD}\| \leq .50$ and stratum group size ≥ 2; utility retention $\geq .95$; over-refusal $\leq .05$; L1 delivery $\geq .95$.
Analysis record	Seed 20260909; Python 3.12.4, NumPy 1.26.4, SciPy 1.13.1. The GPT-4o API interface is documented by (OpenAI, 2026).
Stored records	Canonical records retain task identities and decision windows; selection records retain grid points, doses, and sample membership; E2 records retain operator delivery specifications and paired target/control outcomes.

C.3 Worked evidence-chain checks

Gemma/IPI has $DI_E = .181$ and passes the PAS equivalence test within the .08 bound at $PAS = .035$; SAG raises PAS by .452 and reduces ASR by 11.10 points versus 1.20 for its random control. Delivery, specificity, and utility checks pass, giving local support. Gemma/MP likewise passes CSS at $HS = .160$, $BR = .960$, and $GD = .030$ with a 10.00-point ASR reduction. Gemma/IH reduces ASR by 4.67 points but lacks the E1 signature because $HS = .379 > .20$, yielding E1/E2 conflict. Qwen/TC enters SPS at $M_{SU}(rec) = -2.015$ and $\Delta_{SPA} = .469$; SPS raises the safe-candidate margin by .650 and lowers ASR by 50.53 points, 44.13 more than its random control. Its balance and guardrails pass, yielding local support.