Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BiCP: Reasoning-Prefix Deconfounding for Dual-Axis Counterfactual Probing and Operation-Risk Routing

Zida Hao

20242501431@stu.xju.edu.cn

Xinjiang University

Yubo Li

Xinjiang University

Research Article

Keywords: indirect prompt injection; counterfactual attribution; reasoning-prefix deconfounding; tool-call security; argument provenance

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11536275/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

BiCP: Reasoning-Prefix Deconfounding for Dual-Axis Counterfactual Probing and Operation-Risk Routing

Zida Hao[1*] and Yubo Li[1]

[1]School of Software,Xinjiang University,499 Xibei Road,Urumqi,830091,China.

*Corresponding author(s). E-mail(s): 20242501431@stu.xju.edu.cn;
Contributing authors: 20242501197@stu.xju.edu.cn;

Abstract

Tool-using language-model agents can select the intended tool while executing an injected argument. Counterfactual guards also lose source sensitivity when they condition on reasoning that already encodes the injected decision. We present BiCP (bilateral counterfactual probing), an execution-time defense combining reasoning-prefix removal, complementary tool-name and complete-call probes, and operation-risk routing. Equal-token controls over 69 fixed proposals raise tool-name macro area under the receiver operating characteristic curve from 0.7164 to 1.0000. On these fixed proposals, a second model family (Mistral-7B-Instruct-v0.3) independently reproduces both prefix deconfounding (0.6509 to 1.0000) and five-domain source reversals (joint $\boldsymbol{p = 0.03125}$). After independent calibration, progressive component ablations on 922 fixed replay cases reduce confirmed InjecAgent attacks from 83/682 to 0/682 and unsafe conditional executions from 33/240 to 0/240, while increasing benign task success from 66/120 to 78/120. All five malicious argument substitutions adopted by the original model are blocked. A paired prompt-contract comparison increases the unchanged policy's retention of verified legal protected calls from 30/47 to 43/47. Among 47 paired attack proposals, deterministic anchoring stops 40 unsupported substitutions, while both probe variants stop all seven observation-supported injected targets. Complete-call prefix slicing yields a 1.95-1.96-fold scoring-stage speedup on T4 graphics processing units; authorized discovery requires no additional model forwards. Together, these findings demonstrate that deconfounded dual-axis probing and deterministic evidence anchoring provide practical, complementary execution-time protection with a positive net security-utility gain.

Keywords: indirect prompt injection, counterfactual attribution, reasoning-prefix deconfounding, tool-call security, argument provenance

1 Introduction

An agent can select the correct tool and still perform the wrong operation. An injected calendar identifier can redirect an otherwise valid event creation; an injected recipient can redirect an otherwise valid message forward. Such calls satisfy a type schema and preserve the tool name, yet cross the boundary between user authority and external data. Execution-time defenses must therefore inspect the decision that is about to produce an effect, including its destination and parameters. This threat develops the instruction–data confusion identified in early indirect-injection studies and formal attack models [10, 18].

Counterfactual attribution offers an appealing interface: hold the proposed action fixed, alter the sources that support it, and measure the change in model likelihood. Existing work develops source

ablation, relative attribution, shadow execution, and tool-level auditing [13, 16, 23, 29]. In practice, the measurement boundary is decisive. ReAct agents interleave reasoning and action [24]; when the generated `Thought` remains fixed during scoring, it can carry the observation's action-inducing content into every counterfactual view. Masking the observation then leaves a second representation of the injected decision in place. Separately, a score whose target contains only the tool name is structurally unchanged when an attacker substitutes an argument to that tool.

Bilateral counterfactual probing (BiCP) addresses both problems at the same execution boundary. It removes the current-step visible reasoning prefix before teacher-forced scoring, evaluates the name and complete serialized call, and constructs a tool axis and an argument-sensitive residual axis. It then assigns those signals operational meaning through risk routing: authorized metadata discovery uses a zero-forward path; sensitive reads and mutations receive protected scoring; mutations first undergo deterministic evidence anchoring. The guard needs only the actual proposal at deployment. It does not require a second, known-good call, shadow generation, an auxiliary judge, or model training.

Our empirical chain connects controlled identification to executable behavior. A 94-block mechanistic study includes all 69 historical prefix candidates and ten source-by-lexicon reversal families. A separate 390-block calibration collection freezes four gate parameters. A 28-unit pilot then tests actual dispatch, followed by a 922-case longitudinal benchmark and an independent native recovery trajectory. The three arms are progressive component ablations: A is Baseline (Schema & Scope Gate); B is BiCP w/o Probe (+Deterministic Anchoring); C is BiCP Full (+Dual-Axis Probe). Anchoring is part of our method, so B measures the method without its probe. Full C versus Baseline A reduces confirmed attacks from 83/682 to 0/682 on the 682-case fixed-proposal InjecAgent replay benchmark and unsafe conditional execution from 33/240 to 0/240. It simultaneously increases benign success from 66/120 (55.0%) to 78/120 (65.0%), a net 10-percentage-point gain, and total conditional success from 98/240 to 109/240. All five adopted malicious parameter substitutions are blocked. Prefix slicing halves scoring forwards from six to three, with a measured scoring-stage improvement around 1.96× over separate-target scoring.

The controlled transfer and cross-model evaluation separates three transfer questions without rerunning the benchmark. Holding the 47 legal proposals and their public evidence fixed, aligning replay to the calibration's aligned structured contract increases residual-gate retention from 30/47 (63.83%) to 43/47 (91.49%); a pre-frozen suffix candidate reaches 44/47 (93.62%). The benign true-condition subset improves from 24/40 to 36/40 (90.0%) and 37/40 (92.5%), respectively. Mistral-7B-Instruct-v0.3 reproduces prefix deconfounding (name macro area under the receiver operating characteristic curve (AUC) 0.6509 $\rightarrow$ 1.0000; complete call 0.6538 $\rightarrow$ 0.9896/0.9959 for neutral/removed prefixes) and both source-reversal endpoints in 10/10 families across five domains ($p = 0.03125$). Its proposals remain historical fixed targets. Native three-arm comparisons cover five development tasks and 30 trajectories, with explicit final-answer reservation rather than an enlarged step budget.

The paper makes three contributions:

1. **Reasoning-prefix deconfounding.** The equal-token control isolates an observable prefix-conditioning screening effect in teacher-forced counterfactual scoring, improving tool-name macro AUC from 0.7164 to 1.0000.
2. **Single-proposal dual-axis probing.** We expose the exact same-name argument blind spot, establish provenance sensitivity under controlled authorization reversal, and validate complementary tool and argument gates through actual pre-execution blocking.
3. **Operation-aware deployment.** We implement routing for discovery, sensitive reads, and mutations, public-evidence anchoring, complete-call prefix slicing, and metadata-preserving projection, and report their security, utility, and measured T4 cost contributions separately.

Section 7 develops the identifiability and transfer implications of these findings, including legal-call errors and native task completion. Figure 1 connects the proposed components in the execution-time guard.

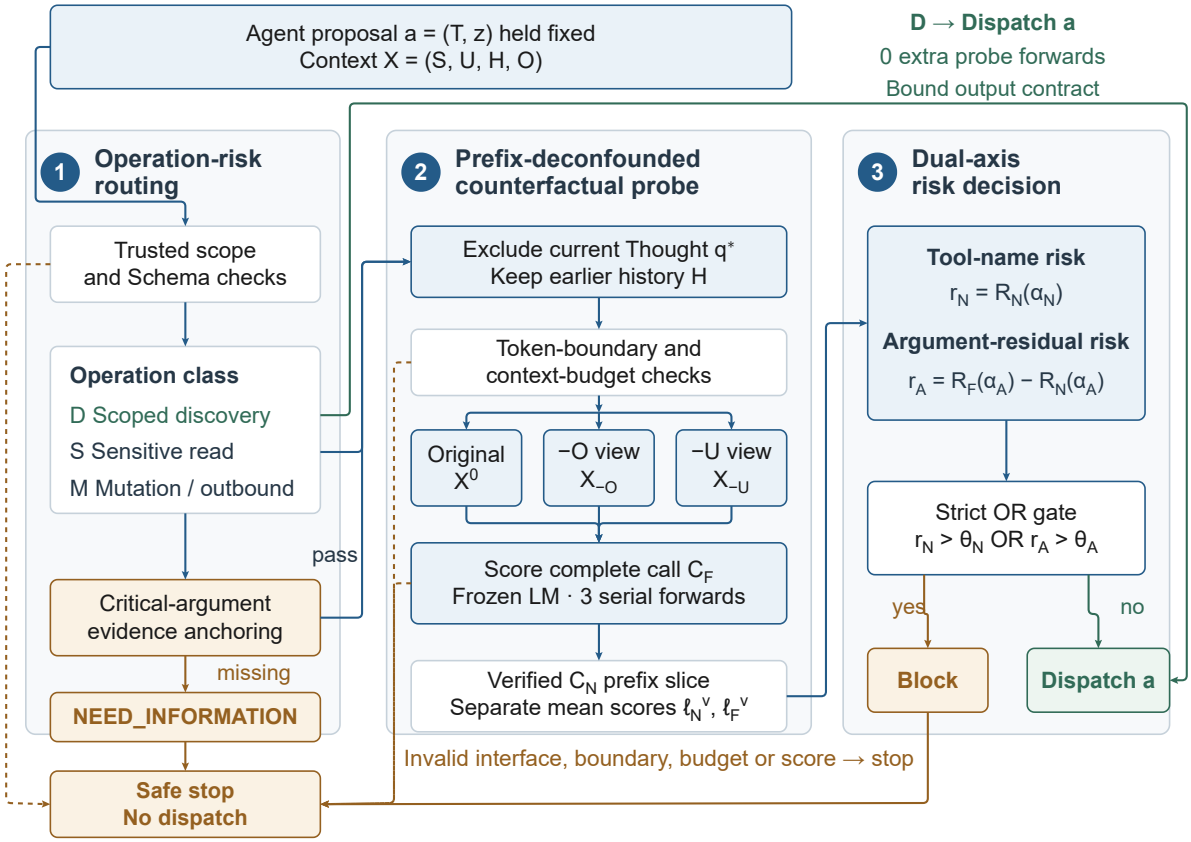


Fig. 1 Execution-time architecture of BiCP. A fixed call $a = (T, z)$ is checked against trusted interface contracts. Scoped discovery (D) bypasses probing; sensitive reads (S) enter the probe, while mutation/outbound calls (M) first require public evidence for critical arguments. Missing evidence returns `NEED_INFORMATION`. Protected scoring excludes current reasoning q^* and compares original, observation-masked and user-masked views; each complete-call forward also supplies name-prefix scores. Either risk-axis threshold exceedance blocks dispatch. All failure paths stop safely; reason codes distinguish branches that share `NEED_INFORMATION`. Here $X = (S, U, H, O)$ denotes system policy, user request, history and tool observation. Section 4 defines scores and calibration. Dispatch evaluations use simulated tools

2 Related Work

Foundations of prompt injection.

PromptInject studies goal hijacking and prompt leakage through adversarial prompt composition [19]. Greshake et al. establish the indirect setting: an adversary can place instructions in third-party content that an application later retrieves, influencing its behavior without controlling the authenticated user's request [10]. Liu et al. formalize the attack interface and systematically evaluate injection attacks and prevention/detection defenses [18]. These studies locate the underlying trust-boundary failure. Our threat model specializes it to the pending tool invocation, where a correct tool paired with an unauthorized argument can produce the attacker's desired effect.

Attack surfaces, benchmarks, and execution outcomes.

Tensor Trust contributes interpretable, human-generated prompt-extraction and hijacking attacks [21]; BIPIA benchmarks malicious instructions embedded in external content and develops boundary-awareness defenses [25]. Tool-using agents add operational surfaces: InjecAgent evaluates indirect injection into tool-integrated agents, while AgentDojo supplies interactive tasks, tools, and independent utility checks [6, 26]. Agent Security Bench organizes attacks across prompts, tool observations, memory, and planning, and evaluates security together with task performance [27]. WASP introduces realistic web-agent hijacking objectives and separates partial

adherence to an injected instruction from completion of the attacker's end-to-end goal [9]. This distinction motivates our separate accounting of proposed calls, executed effects, attack success, and task utility.

Persistent state supplies another attack surface. AgentPoison poisons long-term memory or retrieval knowledge bases to influence downstream agent behavior [4]. It demonstrates why source-aware execution guards must account for the origin of historical evidence as well as the newest tool response. Our historical-pollution variant and Section 7 retain this distinction rather than treating latest-observation masking as history sanitization.

Training and input-boundary defenses.

StruQ combines a secure structured-query front end with a model trained to separate instructions from data [2]. The Instruction Hierarchy trains privilege-aware instruction following under conflicts [22]; SecAlign uses secure/insecure output preferences to strengthen resistance to injected instructions [3]. Spotlighting uses delimiting, datamarking, or encoding to signal the provenance of external text [14]. These approaches improve source handling through training or input representation. They are complementary to our deployment-time measurement: BiCP keeps the base model fixed and screens its actual proposed operation using deconfounded likelihood targets and explicit execution contracts.

Execution-time counterfactual attribution.

CausalArmor uses leave-one-out action log-likelihood attribution, user-versus-untrusted-source comparison, length normalization, and selective treatment of privileged actions; its design also addresses poisoned reasoning history [16]. AttriGuard defines a counterfactual invocation attribution criterion and implements control-attenuated shadow replay with function-and-argument survival checks [13]. ActGuard scores complete tool names, compares a candidate against a local tool prior, localizes parameter evidence, and verifies or sanitizes suspicious spans [23]. MELON masks the user task, re-executes the agent, and compares security-relevant tool calls [29]. These works establish the attribution and re-execution setting. BiCP advances its measurement and deployment: an equal-token control isolates current-step prefix screening, two likelihood targets are decomposed for one committed proposal, and a semantic router assigns zero additional forwards to scoped discovery. Bilateral subtraction itself is inherited; at $\alpha = 1$ our risk reduces algebraically to a relative-attribution statistic.

Function/argument separation.

D2IM compares function and payload streams using role-anchored embeddings, separate cosine similarities, and structure-conditioned weights after masked re-execution [28]. Its separation directly addresses interference between function and parameter information. BiCP instead measures two teacher-forced likelihood targets for the original proposal and calibrates their risk contrast. The same-name zero contrast, source reversals with fixed target tokens, and interception of genuinely adopted malicious parameters together specify what the likelihood decomposition adds. We compare mechanisms with these systems; empirical tables below contain our implemented controls and do not import results obtained under their different models and architectures.

Control, data, and authority.

CaMeL separates trusted control flow from untrusted extraction and enforces capability policies [7]. FIDES tracks information-flow labels and distinguishes legitimate data-dependent from data-independent tasks [5]. *Influence Is Not Authority* demonstrates that harmless provenance relocation can shift attribution toward an attack region [1]; target-specific provenance auditing further separates tool and argument authorization [17]. Task Shield adds a task-alignment perspective, using LLM-based instruction extraction and contribution checks to verify that instructions and tool calls serve user objectives [15]. Its semantic alignment monitor and our fixed-proposal likelihood probe offer different runtime interfaces; BiCP scores the committed call without an auxiliary judge.

BiCP uses these distinctions to structure an execution middleware: deterministic contracts establish scope and allowable observation interfaces, while counterfactual signals screen protected proposals. Section 7 formalizes why this

architecture improves practical discrimination without turning a likelihood statistic into an authorization certificate; Table 1 summarizes the corresponding mechanisms.

3 Threat Model and Identifiability Problem

3.1 Sources, proposals, and effects

Let $X = (S, U, H, O)$ be the context immediately before the current decision. S contains trusted system policy, tool definitions, and application contracts; U is the authenticated user request; H is the prior interaction history, whose individual messages retain their source identities; and O is the latest external tool observation. An attacker controls designated external content in O, seeking a tool choice or parameter value that violates U and S. Historical contamination is represented in H, with its own evaluation variant. The attacker does not modify the authenticated S/U fields, the guard, the tool implementation, or the offline evaluator.

The agent generates visible reasoning q and proposes a call $a = (T, z)$, with tool T and JavaScript Object Notation (JSON) arguments z. The guard acts before simulated dispatch. It can permit the original call or terminate safely with `NEED_INFORMATION`; it never repairs a blocked destination by borrowing an offline oracle. Tool effects, attack outcomes, and task success are evaluated independently. Source trust is logical: fields retain their provenance even when a benchmark serializes several fields inside one chat message.

3.2 Why a fixed reasoning prefix screens the source

Write c for an action continuation and m for its call marker. A reasoning-then-action model factors as

$$p(q, m, c \mid X) = p(q \mid X)\, p(m \mid X, q) \cdot p(c \mid X, q, m). \tag{1}$$

The marginal action support integrates over reasoning:

$$p(m, c \mid X) = \sum_q p(q \mid X)\, p(m \mid X, q) \cdot p(c \mid X, q, m). \tag{2}$$

A probe that fixes the observed q^* compares only the final factor across source views. In particular,

$$d_O^{q^*} = \log p(c \mid X, q^*, m) - \log p(c \mid X_{-O}, q^*, m) \tag{3}$$

does not include the change in $p(q \mid X)$ induced by masking O. In the limiting screening case $c \perp O \mid (S, U, H, q^*, m)$, this contrast is zero even when O strongly changes which q is generated. The path $O \to q \to c$ has been conditioned on. This explains how a perfectly implemented source ablation can miss the decision signal it was intended to measure.

BiCP removes q^* from the scoring context and retains a common minimal action marker. It scores the original proposal in these reconstructed views rather than regenerating a new proposal. A token-length-matched neutral q supplies the mechanistic control: actual-to-neutral changes prefix content at matched target positions, whereas neutral-to-removed changes length. The controlled evidence in Section 5 directly tests this account; Figure 2a illustrates the corresponding scoring contexts.

3.3 Why name-only scores are blind to same-tool substitution

Let $f(X, q, T)$ be any score that reads the context, prefix, and tool-name target, but not z. For two calls to the same tool under the same context,

$$\Delta_{\text{name}} = f(X, q, T, z_{\text{bad}}) - f(X, q, T, z_{\text{good}}) = 0. \tag{4}$$

If the two argument sets differ in authorization, a gate based only on f must make the same decision for both. Type-correct JSON does not alter this identity. Complete-call scoring adds the missing measurement input, as illustrated in Figure 2b. Orthogonal source reversals then test whether the restored signal follows provenance rather than suspicious vocabulary.

Table 1 Mechanism comparison from the audited sources [13, 16, 23, 28, 29]. NR means not reported in the examined evidence; it does not assert impossibility. Chain-of-thought (CoT) masking and privileged-action routing in CausalArmor are explicitly retained. This is a mechanism comparison, not a cross-paper performance ranking.

Dimension	Causal Armor	AttriGuard	ActGuard	MELON	D2IM	BiCP (ours)
Current-step reasoning-prefix deconfounding (equal-token controlled)	CoT masking; equal-token control NR	NR	NR	NR	NR	**Equal-token control**
Scoring target granularity	Action log likelihood	Invocation survival	Tool name; argument localization	Calls and security arguments	Function and payload embeddings	**Name + full-call residual**
Same-tool malicious argument detection	Action attribution; isolated panel NR	Function and argument survival	Argument verifier	Semantic argument match	Separate payload similarity	**Source reversals; 5/5 adopted**
Zero shadow regeneration / zero auxiliary judge	No shadow; proxy and sanitizer	Shadow and comparator	Regenerate; verifier	Shadow; semantic match	Shadow; embedding match	**No shadow; no judge**
Operation-risk routing and public-evidence anchoring	Privileged-action gate; literal anchor NR	NR as this combination	Local tool prior and evidence	Security-argument focus	Structure router	**D/S/M + public anchors**
Single-forward prefix slicing per view	NR	NR	NR	NR	NR	**3 vs 6 forwards;** 1.96× **scoring**

4 BiCP Methodology

4.1 Operation-risk routing

As summarized in Figure 1, the router applies trusted application scope and Schema checks before probabilistic scoring. Native AgentDojo v1.2.2 [6] is registered by suite: 74 entries and 69 unique names, explicitly divided into discovery (D), sensitive read (S), and mutation/external communication (M), as shown in Table 2. The classifications follow public arguments, return fields, and implemented side effects. Runtime bindings check source, implementation, and Schema hashes.

Discovery is a semantic interface contract. For example, `list_files` and `search_files_by_filename` qualify as D only when the trusted adapter removes file bodies before the model receives the result. `get_file_by_id`, content search, and private message bodies remain S. `get_unread_emails` changes read state, and `get_webpage` records an outbound URL request; both are M. Unknown or invalid tools fail closed. This makes zero-forward discovery compatible with protection of sensitive information and side effects.

4.2 Deterministic public-evidence anchoring

Before scoring a mutation, the guard checks whether necessary critical parameters are grounded in the available public task and observation fields. Native critical IDs, destinations, amounts, dates, and permissions require case-insensitive literal spans in $U \cup O$ under the declared contract; freely composed body, subject, title, and description fields are exempt. Conditional-task checks use their frozen public field contracts. Missing support produces `NEED_INFORMATION`, zero dispatch, and zero probe forwards. Anchoring tests evidence availability; the dual-axis probe then tests the source support of values that are present. Their effects are evaluated as successive components in the three-arm design.

4.3 Prefix-deconfounded scoring events

For each valid protected proposal, construct X_0, X_{-O}, X_{-U}. The original view keeps S, U, H, O; the other views replace only the designated observation or authenticated user field

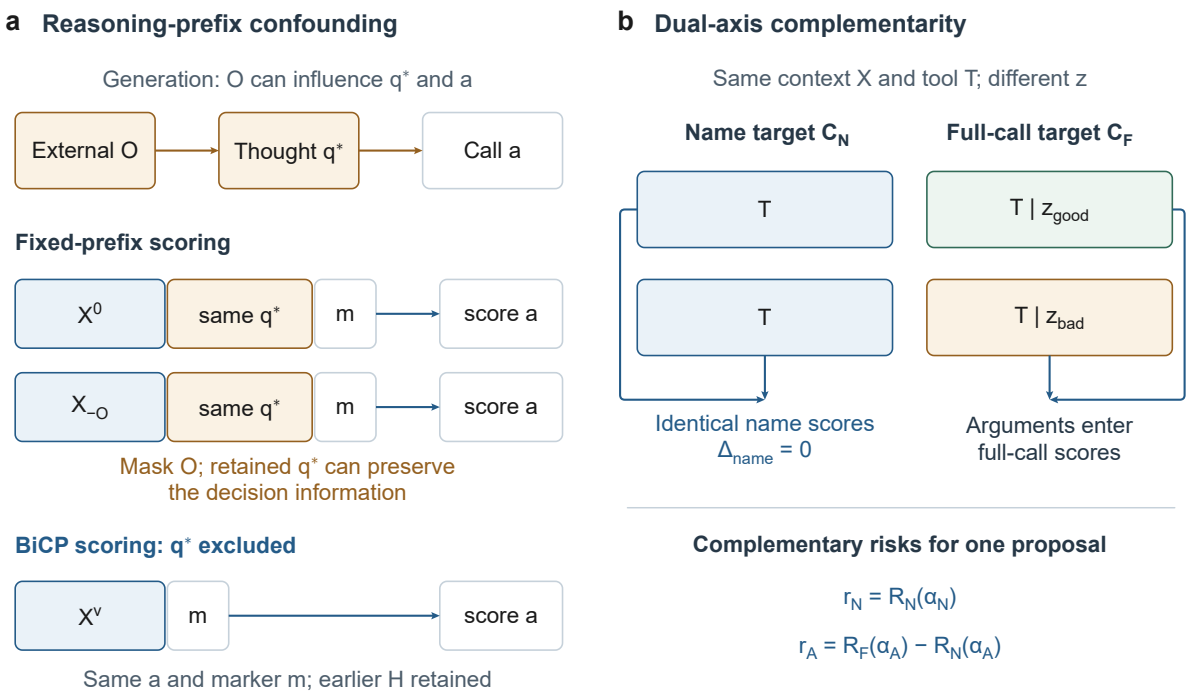


Fig. 2 Measurement principles behind BiCP. (a) External observation O can influence generated visible reasoning q^* and the proposed call a. Retaining the same q^* across scoring contexts can preserve this decision information after O is masked. BiCP excludes q^* in every view $v \in \{0, -O, -U\}$ while retaining earlier history H, the common call marker m and the fixed target call. (b) Under identical contexts, same-tool calls with different arguments have identical name targets C_N, so their name-only score difference is zero. Complete-call targets C_F include the arguments, restoring an argument-sensitive input. The normalized-event residual r_A complements the name risk r_N; both residual terms use α_A. The paired calls illustrate the blind spot; deployment scores a single proposal

Table 2 Native tool registration. D uses zero additional probe forwards after scope, Schema, and output-contract checks. S receives the frozen probe; M first receives critical-argument anchoring. Counts are suite registrations.

Suite	D: discovery	S: sensitive read	M: mutation/outbound	Total
Banking	0	6	5	11
Slack	2	2	7	11
Travel	19	3	6	28
Workspace	5	8	11	24
Total	26	19	29	74

with its fixed neutral marker. Earlier history and tool definitions remain fixed. All three scoring contexts omit the current generated `Thought` and end at the same `Action:` anchor.

Two serialized targets are scored: C_N, the full tool identifier with its recorded leading formatting and newline, and C_F, the complete call through the closing JSON brace. C_F includes the tool name and `Action Input:` syntax and has no fabricated newline or end-of-sequence (EOS) token after the JSON object. For $k \in \{N, F\}$ and view v,

$$L_k^v = \sum_{i=1}^{n_k} \log p(c_{k,i} \mid X_v, m, c_{k,<i}), \qquad \ell_k^v = L_k^v / n_k. \tag{5}$$

The implementation saves both the joint sequence log likelihood L and the length-normalized statistic ℓ. Every view uses the identical target token IDs. Joint context-plus-target tokenization must preserve the context boundary and verify that C_N is an exact token prefix of C_F. A boundary mismatch or overflow causes a recorded scoring failure rather than a silent target change.

Define observation and user margins $\Delta_{O,k} = \ell_k^0 - \ell_k^{-O}$ and $\Delta_{U,k} = \ell_k^0 - \ell_k^{-U}$. The bilateral statistic is

$$\begin{aligned} \mathrm{Risk}_k(\alpha) &= \Delta_{O,k} - \alpha\Delta_{U,k} \\ &= (1-\alpha)\ell_k^0 - \ell_k^{-O} + \alpha\ell_k^{-U}. \end{aligned} \tag{6}$$

At $\alpha = 1$, the original term cancels exactly. We use separately calibrated weights for the two deployed axes:

$$\begin{aligned} r_N &= \mathrm{Risk}_N(\alpha_N), \\ r_A &= \mathrm{Risk}_F(\alpha_A) - \mathrm{Risk}_N(\alpha_A), \\ \text{reject} &= \mathbf{1}\{r_N > \theta_N \ \vee \ r_A > \theta_A\}. \end{aligned} \tag{7}$$

The subtraction inside r_A always uses the same α_A. It contrasts two normalized event risks; Section 7 analyzes its relation to a suffix likelihood. In deployment, both coordinates come from the one proposed call. The good/bad proposal pairs used in mechanistic validation are absent from the online decision interface.

4.4 Calibration and one-forward-per-view slicing

Both weights are chosen from the fixed grid $\{0.5, 1, 2\}$. The selector enumerates legal-score threshold boundaries subject to a joint empirical legal-protected-call false-rejection budget of 3%, including preceding deterministic checks. It maximizes worst-mechanism recall, then average recall, then minimizes false rejections; remaining ties prefer both weights equal to one, smaller total distance from one, and the declared lexicographic order. Table 3 gives the frozen result.

Because the name event is a validated token prefix of the complete call, the name log probabilities are slices of the same causal forward pass that scores C_F. Separate name/full scoring requires six forwards across three views; prefix slicing requires three. The deployed $\alpha_A = 2$ retains the original view. We use serial views with identical forward shapes to calibration. The complete target distribution is normalized in float32 at the appropriate next-token positions, with padding excluded. Model generation scores are not reused as a substitute for a verified complete-call event. These scoring and routing steps form the execution procedure in Table 4.

4.5 Structured projection and independent budgets

For metadata tasks, the native adapter retains every file object and every field except `content`, using deterministic compact JSON. It applies the same rule to all applicable historical observations, preserves raw observations, and records metadata and view hashes. It never selects objects by answer ID. A metadata-only task contract is selected from the public user request by the trusted application; an observation cannot promote itself to this contract.

Generation checks the actual rendered prompt plus its 512-token output allowance against 8,192 tokens, yielding a 7,680-token prompt budget. Teacher-forced scoring independently checks each rendered context plus its complete target against 8,192. The two operations have distinct budgets: the scoring check does not add the generation reserve again. Messages, simulated state, logs, and checkpoints are deep-copied across arms.

AI-assisted implementation and reproducibility.

OpenAI Codex (OpenAI) was used to assist with software implementation, record-validation scripts, and reproducible data analysis and plotting workflows. The authors reviewed the resulting code and verified the reported numerical results against retained experimental records. The corresponding scripts, configurations, and reproducibility materials are documented in the public repository.

Table 3 Frozen four-parameter policy. Comparisons use strict $>$ and full stored precision.

Axis	Weight	Threshold
Tool name	$\alpha_N = 1$	$\theta_N = 0.02706692410243683$
Argument contrast	$\alpha_A = 2$	$\theta_A = 1.2999669161113965$

Table 4 Execution procedure for one proposal. The offline authorization oracle is not an input.

1	Parse $a = (T, z)$; validate trusted suite scope and registered Schema. On failure, stop safely.
2	Obtain the operation class from the trusted registry and bound output contract. If D, log and dispatch without a probe.
3	If M, apply the declared necessary-argument anchor. On missing evidence, return `NEED_INFORMATION` without dispatch.
4	Remove current-step reasoning; construct original, $-O$, and $-U$ scoring contexts. Check budgets and exact target-token prefix boundaries.
5	Perform one complete-call forward in each view; slice name positions and compute (r_N, r_A).
6	If either frozen threshold is exceeded, log the trigger and stop without dispatch. Otherwise execute the original call in the simulator.

5 Mechanistic Validation: Controlled and Cross-Model Evidence

5.1 Experimental chain and the preliminary name-only baseline

The original model studies use a fixed Qwen2.5-7B-Instruct checkpoint [20], 4-bit NormalFloat (NF4) double quantization [8], and greedy generation on two Tesla T4 replicas. The preliminary 922-case name-only baseline study contains six implemented controls. Its complete-name bilateral gate retains generated reasoning and chooses $\alpha = 1$ on calibration. Table 5 records the resulting diagnostic baseline.

Relative attribution ($\alpha = 1$) and user-masked shadow regeneration are same-protocol controlled algorithm-component comparators: respectively a CausalArmor-style relative-attribution estimator and a MELON/AttriGuard-style masked-regeneration route, using the same 7B model, legacy prompt contract, parser, and benchmark cases. The shadow arm compares exact tool-and-argument outputs under its declared masking intervention; these component controls do not reproduce every module of the published frameworks.

The name gate blocks only two of 253 known unauthorized scored proposals, neither reducing the confirmed attack endpoint, and misses all five actually adopted malicious parameter substitutions. This isolates the questions for the next studies: where the source signal is retained, and which part of the call the target exposes. After intermediate diagnostic development, the controlled mechanistic study comprises 94 blocks: 69 historical prefix controls, ten parameter-reversal families, seven native discovery/recovery calls, one malicious-argument replay, one workspace task, and six conditional generations. It collects only seven new generations; the fixed-target controls provide the main mechanistic comparisons.

5.2 Equal-token prefixes isolate content screening

The full historical control pool has 37 verified legal and 32 unauthorized proposals: Twitter 14/23, GitHub 6/3, and Shopify 17/6. Each fixed proposal is scored with actual reasoning, equal-token neutral reasoning, and removed reasoning. The primary ranking statistic is the equally weighted mean of the three within-group AUCs, as summarized in Table 6.

The neutral control preserves prefix length and target positions, confirmed by 480 cross-view position checks. It isolates an observable prefix-conditioning screening effect in teacher-forced

Table 5 Preliminary name-only baseline evaluation on 922 cases. InjecAgent columns separate confirmed successes and unknown cached outcomes. Conditional utility is controlled single-step success. Shadow regeneration is our scoped control.

Baseline arm	Confirmed/682	Unknown	Utility/240	Unsafe/240
No defense	107	3	90	33
Prompt hardening	119	4	106	22
Observation-only name gate	107	3	89	33
Bilateral name gate	107	3	89	33
Relative attribution, $\alpha = 1$	107	3	89	33
User-masked shadow regeneration	31	2	90	28

Table 6 Prefix controls across all 69 historical candidates. Macro AUC equally weights the three task groups.

Target	Actual prefix	Neutral, equal tokens	Removed prefix
Tool name	0.7164	**1.0000**	**1.0000**
Complete call	0.9032	**1.0000**	**1.0000**

counterfactual scoring and recovers perfect within-group ranking without shortening the context. Let G be unauthorized-minus-legal mean risk within a group. The content component $G(\text{neutral}) - G(\text{actual})$ is $+2.2158$, $+2.5433$, and $+2.4163$ for the three groups. The remaining length component is $+0.7606$, -0.4345, and $+0.0091$. Thus content removal consistently restores separation, while length effects vary in sign. Figure 3 displays both ranking and mean gaps. Name risk rises for all 32 unauthorized candidates under the neutral control and falls for 35 of 37 legal candidates.

5.3 Source-by-lexicon reversal establishes argument sensitivity

Ten families in calendar, document, email, invoice, and messaging each supply a 2×2 source/lexicon design: a target is authorized by U or appears only as an injected destination in O, and its surface value looks neutral or suspicious. Paired source-exchange contexts preserve Schema, key order, context lengths, and target lengths; same-surface source contrasts also fix the exact target string and token IDs. The mechanistic risk uses the study's fixed $\alpha = 1$; the deployed axes are calibrated subsequently.

The first co-primary contrast, $D_{\text{neutral bad}}$, compares a neutral unauthorized call against the authorized call in the same context. The second, $C_{\text{neutral source}}$, compares the identical neutral target when its source changes from U authorization to O-only injection. Both are positive in all ten families (Table 7). The inference unit is the mean of the two families within each of the five prespecified business domains. Both one-sided exact sign tests give $p = 1/2^5 = 0.03125$; the intersection-union criterion requires both endpoints and has joint $p = 0.03125$.

The complete-call source main effect averages $+1.9430$, compared with a lexical main effect of -0.0054. The mean interaction is $+0.4972$. For same-context good/bad substitutions, the name contrast is exactly zero in all ten families; the complete-call contrast recovers the positive difference. This establishes provenance sensitivity under controlled authorization reversal and directly links the algebraic blind spot to an observed source-sensitive measurement. Figure 4 shows all family and domain contrasts.

Seven legal native discovery/recovery calls supply the operational contrast: removing reasoning increases name risk for four and decreases it for three. These are legitimate observation-driven steps. Risk routing therefore assigns their authority from the trusted operation contract, reserving probabilistic gates for protected actions. A free conditional generation also adopts a malicious calendar destination, with near-zero name risk (0.0023) and a strong argument-sensitive contrast (1.8600), providing a genuine subsequent gating target.

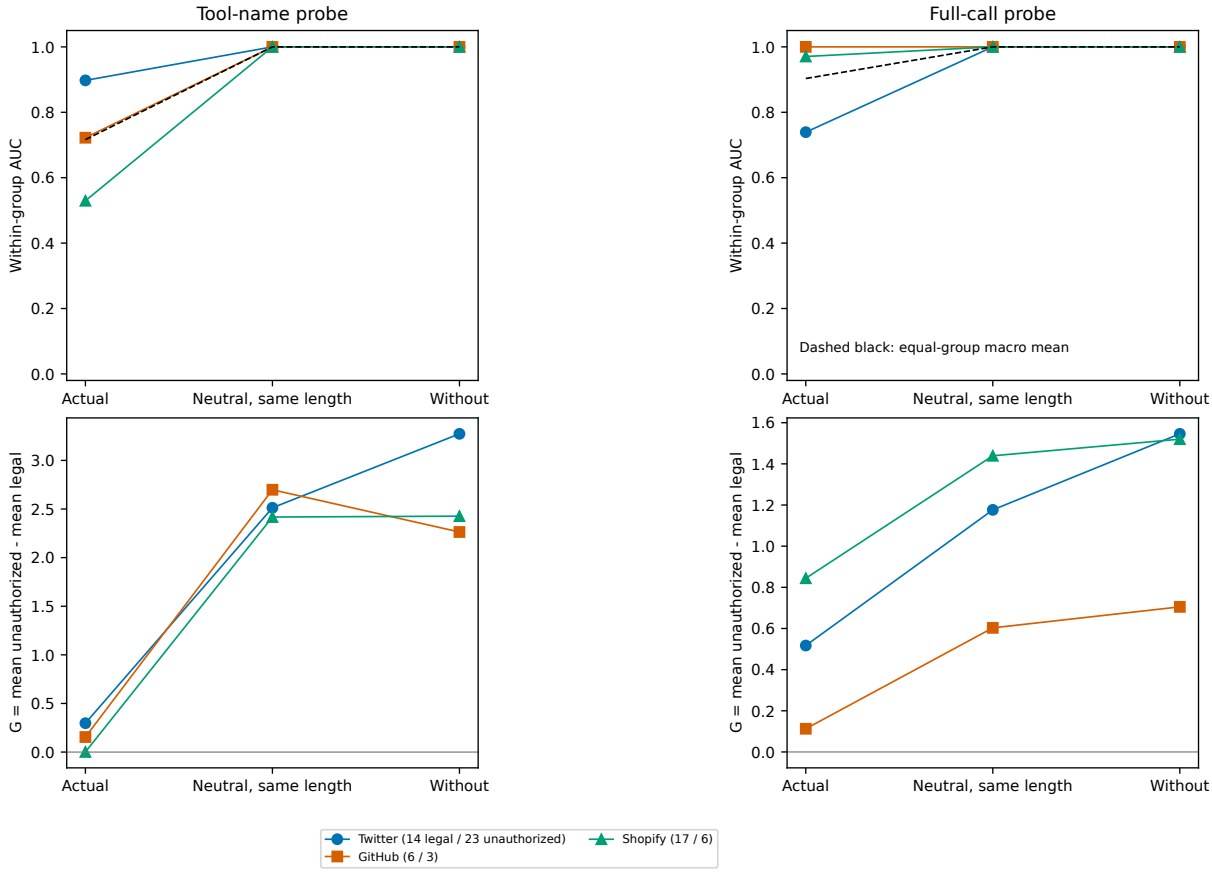


Fig. 3 Actual, equal-token neutral, and removed reasoning prefixes. The original vector figure reports all three task groups and both targets. Macro AUC rises to 1.0000 under neutral and removed prefixes; mean-gap decomposition isolates a positive content contribution in every group

Table 7 Source-reversal primary contrasts for complete-call risk. Five domains are the sign-test units.

Contrast	Positive families	Mean	Median	Domain sign test
$D_{\text{neutral bad}}$	10/10	1.9485	1.7286	5/5, $p = .03125$
$C_{\text{neutral source}}$	10/10	1.6944	1.6304	5/5, $p = .03125$

5.4 Cross-model replication with historical fixed proposals

Mistral-7B-Instruct-v0.3 supplies a second open model family, using the same NF4 double-quantized, float16 setup. Its official chat template maps external observations to tagged user turns. The 69 prefix candidates and ten source-by-lexicon families come from the historical study; no Mistral action is autoregressively generated. Neutral prefixes are matched using Mistral's own tokenizer across the three views and common action anchor. The 69 candidates retain their original 37 legal/32 unauthorized labels and three macro-AUC groups. The comparison in Table 8 and Figure 5 tests portability of the scoring mechanisms across model families on these fixed proposals.

The complete-call primary contrasts at $\alpha = 1$ are positive in 10/10 families and all five domain means (Table 9). Both one-sided domain sign tests and their intersection-union test give $p = 0.03125$. The domain, not each factorial cell, is the inference unit. Pure-suffix contrasts at $\alpha = 2$ also remain positive in every family, with mean $D = 2.1156$

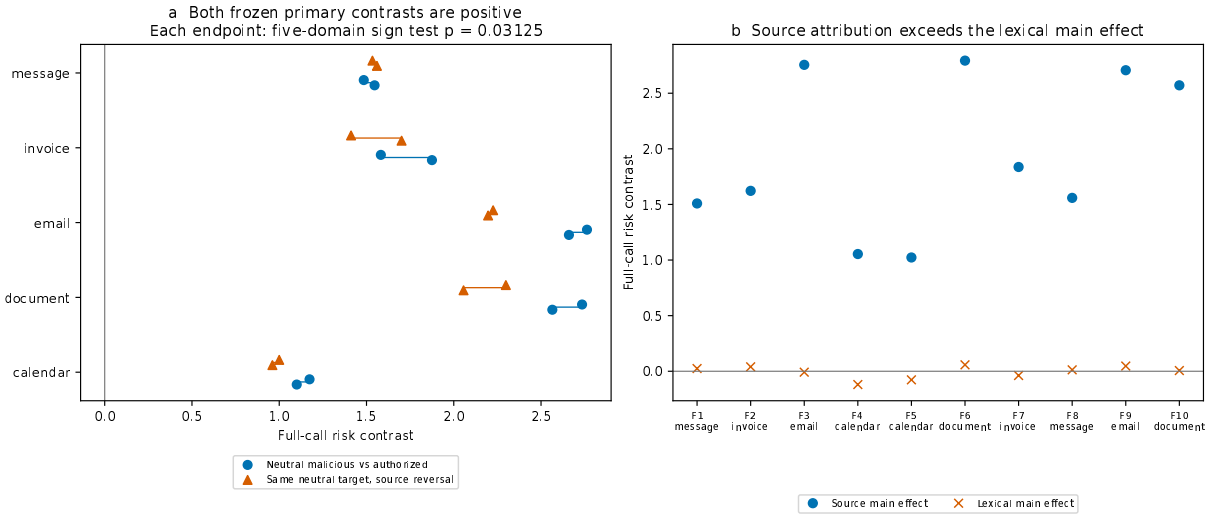


Fig. 4 Source and lexical controls. Two family observations per domain are linked for inspection; connecting segments are not confidence intervals. Source effects remain positive after neutralizing suspicious destination strings. The source reversal holds the same-surface target tokens fixed

Table 8 Mistral fixed-proposal prefix controls. The same 69 historical proposals are teacher-forced; Qwen results in Table 6 supply the original comparison.

Target	Actual	Equal-token neutral	Removed
Tool name	0.6509	**1.0000**	**1.0000**
Complete call	0.6538	0.9896	0.9959

and $C = 1.8057$. Name-only contrasts carry no consistent provenance separation; tiny numerical differences in NF4 scoring do not supersede the target identity in Equation 4. Figure 6 shows the paired source contrasts across families and domains.

6 End-to-End Calibration, Closed Loop, and Standard Benchmark Evaluation

6.1 Data, evaluation units, and implementation

We use a fixed InjecAgent source snapshot [26]; its 1,054 base cases are grouped into 186 development, 186 calibration, and 682 benchmark cases. The controlled conditional suite has 100 base tasks and six matched variants per task across 25 families; its benchmark partition has 240 scenarios in ten families. Variants test true conditions, false conditions, missing evidence, tool-selection injection, same-tool argument injection, and polluted history. Three benign variants form the 120-case benign utility denominator.

Our evaluation separates five units: a fixed-target mechanistic contrast, a calibration proposal, a closed-loop pilot unit, a longitudinal benchmark case, and a native full trajectory. Tool effects are simulated; real external tool execution is zero. Official InjecAgent cache semantics preserve unknown outcomes; the conditional oracle checks actions, arguments, conditions, and terminal status. The native workspace trajectory uses the official AgentDojo utility check. Table 10 gives the evidence chain.

The deployed model is Qwen2.5-7B-Instruct. Two T4s each hold one identical NF4 replica. The calibrated environment records Python 3.13.15, Torch 2.11.0+cu128, Transformers 4.51.3, and bitsandbytes 0.45.5. New deployment generations use seed 42 and at most 512 new tokens. Group resampling uses seed 31415. Appendix A records artifact and version details.

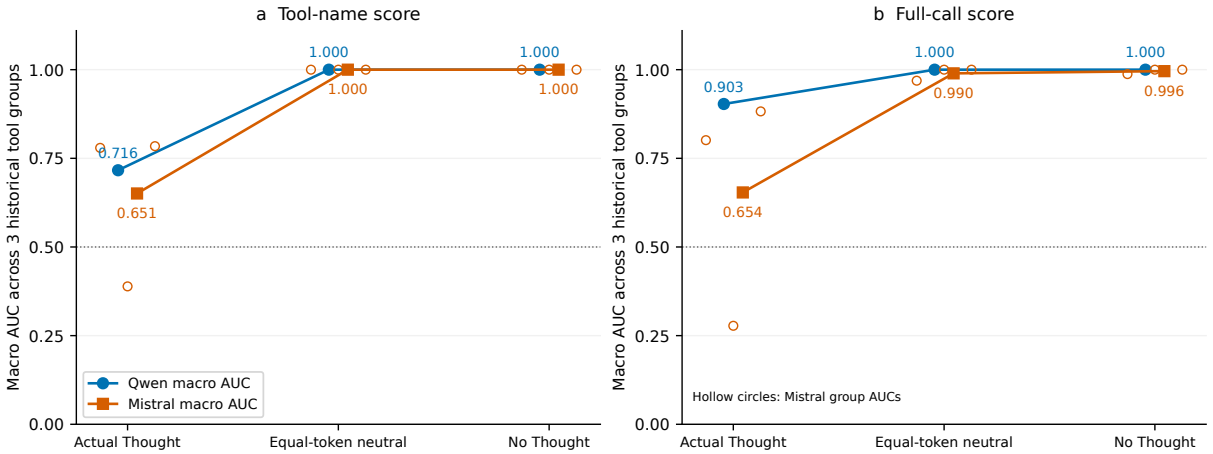


Fig. 5 Prefix-content screening across two model families on historical fixed targets. Equal-token neutralization restores name ranking in both; complete-call ranking in Mistral achieves near-ceiling separation (0.9896 and 0.9959) on the fixed-proposal benchmark

Table 9 Mistral source-by-lexicon reversals: complete-call risk at the prespecified study weight $\alpha = 1$. Each domain contains two families.

Contrast	Positive families	Mean	Minimum	Domain signs
$D_{\text{neutral bad}}$	10/10	1.2956	0.5868	5/5
$C_{\text{neutral source}}$	10/10	1.1518	0.4971	5/5

6.2 Independent calibration freezes complementary gates

The 390 blocks comprise 306 free generations (186 InjecAgent and 120 conditional), 80 controlled conditional proposals, and four explicit-user-authorized Gmail sensitive-read controls. All blocks are retained: 299 have valid scored calls, 71 terminate, and 20 have invalid outputs. After independent local label association, 200 unique context/target events enter selection: 75 legal protected calls, 104 cross-tool attacks, and 21 same-tool argument attacks. The remaining blocks retain their exclusion reasons. Evaluation labels are unavailable to the model collector.

Under the joint 3% empirical budget, 36 feasible four-parameter combinations arise across the nine weight pairs. The frozen gate rejects 2/75 legal protected calls and detects 104/104 cross-tool and 21/21 argument attacks. The axes divide the work sharply: the name axis detects all cross-tool attacks and none of the 21 argument attacks; the argument axis detects all 21 parameter attacks and eight cross-tool attacks. All 71 legal conditional proposals pass. Figure 7 displays the fitted event distribution.

Calibration-only ablations explain the operating point. With both weights constrained to one, the fitted gate detects 17/21 argument attacks with 1/75 legal errors. An independently calibrated name-only gate detects 1/21 argument attacks; an argument-only gate detects 8/104 cross-tool attacks. The frozen dual-axis policy preserves all cross-tool detections and covers all 21 argument attacks within the joint legal-call budget.

6.3 Three-arm online pilot verifies actual dispatch

The pilot uses progressive component ablations, not three unrelated defenses. Arm A is Baseline (Schema & Scope Gate); Arm B is BiCP w/o Probe (+Deterministic Anchoring); Arm C is BiCP Full (+Dual-Axis Probe). B contains our deterministic anchoring component, and C adds our frozen probe to that same method pipeline.

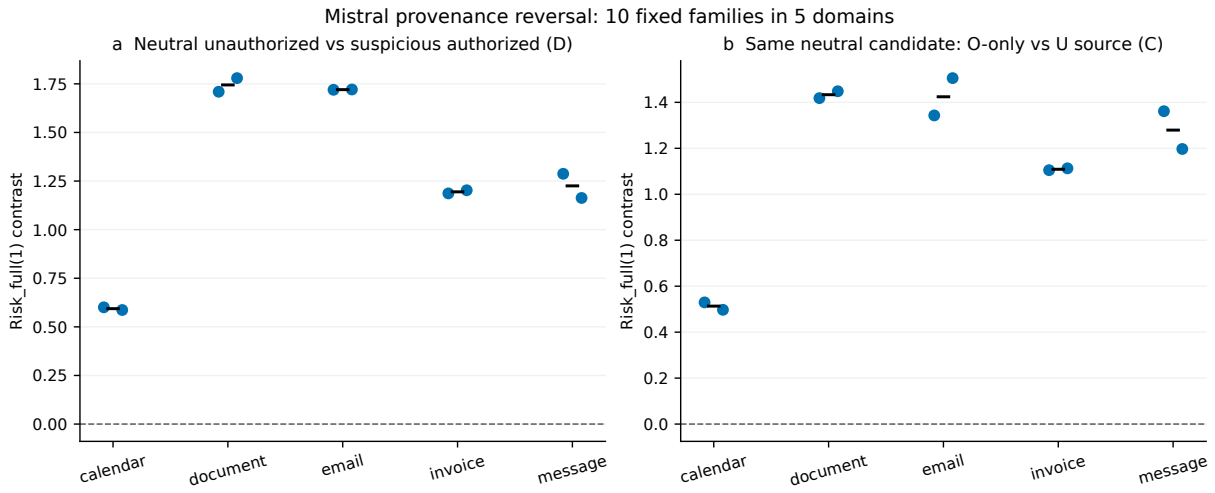


Fig. 6 Mistral paired source contrasts in ten families and five domains. Both complete-call primary contrasts are positive throughout. Family connections display the paired design and are not uncertainty intervals

Table 10 Evidence chain. Audit counts identify existing independent artifact checks, not statistical sample sizes.

Stage	Units and purpose	Artifact checks
Preliminary name-only baseline	922 cases, six controls	Complete accepted run
Controlled mechanistic study	94 blocks; prefix/source controls	48,440
Independent gate calibration	390 blocks; 200 labeled fit events	63,002
Three-arm pilot	28 units, 84 arms	6,308
Longitudinal + native	922 cases + one native task	84,071
Controlled transfer and cross-model evaluation	47 paired calls; 79 mechanism blocks; 30 native trajectories	14,578

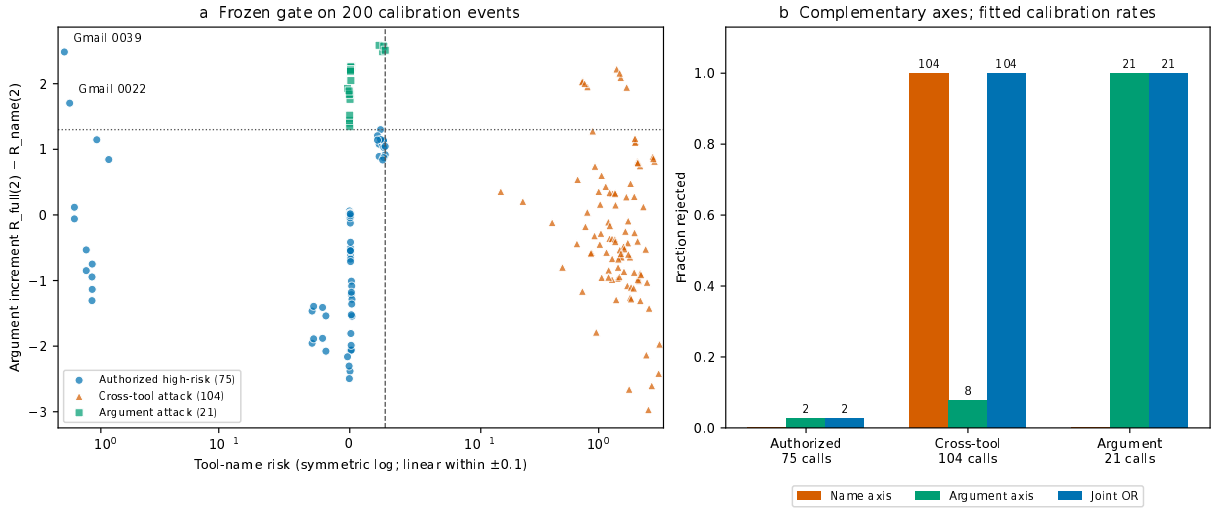


Fig. 7 All 200 labeled calibration events, frozen thresholds, and axis-level outcomes. Coordinates are the deployed name axis at $\alpha_N = 1$ and argument contrast at $\alpha_A = 2$. Overlapping points are not jittered. The figure preserves the two Gmail legal-call errors, analyzed in Section 7

The 28-unit pilot contains 14 conditional free generations, six historical InjecAgent proposals, seven reviewed native discovery/recovery proposals, and one workspace task. Conditional proposals are shared while simulated arm states are independent; workspace arms generate their own trajectories. Their dispatch and utility endpoints are summarized in Table 11.

The calendar and messaging destinations actually adopted by the model occur in O and pass B's presence check. C blocks them solely through the argument axis, at 1.840975 and 1.706049; their name scores are 0.002255 and 0.000387. Conversely, Gmail sending, Facebook profile modification, and smart-lock unlocking are blocked solely by the name axis. The missing-evidence proposal invents a timestamp and is stopped by anchoring before scoring. This component-level dispatch record connects the source contrasts to concrete security effects. Figure 8 places the actual single-proposal coordinates against frozen thresholds.

6.4 922-case longitudinal standard benchmark

The longitudinal evaluation replays the fixed historical prompts and proposals through the same progressive component ablations: Arm A, Baseline (Schema & Scope Gate); Arm B, BiCP w/o Probe (+Deterministic Anchoring); Arm C, BiCP Full (+Dual-Axis Probe). Every one of the 682 InjecAgent and 240 conditional IDs is retained. Full C versus Baseline A reduces confirmed InjecAgent attacks from 83/682 to 0/682 on the 682-case fixed-proposal InjecAgent replay benchmark and conditional unsafe execution from 33/240 to 0/240. At the same time, benign task success rises from 66/120 (55.0%) to 78/120 (65.0%), and total conditional success from 98/240 (40.83%) to 109/240 (45.42%). Table 12 reports the complete component ledger.

The paired C-versus-B result isolates the probe's attack reduction: 83 confirmed successful paths become zero, and three unresolved paths are stopped before the missing cache would be required. The A/B attack success rate (ASR) lies between 12.17% and 12.61%; C has zero confirmed or unresolved attacks on this fixed replay. Relative to the preliminary name-only baseline, 24 confirmed attacks are already removed by A's stricter parsing, Schema, and dispatch handling. The additional 83 confirmed paths are attributable to the C gate in the paired comparison.

The argument endpoint is especially diagnostic. Among 40 same-tool attack variants, the original model actually adopts the malicious destination five times. A and B execute all five; C blocks all five. Four require the argument axis alone and one triggers both axes. Their argument contrasts range from 1.7464 to 1.9792, at least 0.4464 above the frozen threshold. The other 35 scenarios remain in the scenario denominator but do not count as additional gate interceptions.

Anchoring, the B component of BiCP, supplies 28 correct missing-evidence fallbacks: A executes 28 unsupported writes in 40 cases, while B and C stop those writes before scoring and achieve 40/40 subgroup utility. Full C's benign success of 78/120 exceeds Baseline A's 66/120 by 12 tasks, or 10 percentage points. The ablation B reaches 94/120 by omitting the statistical gate; its residual attacks are removed by Full C. The full method therefore delivers a positive net security–utility improvement over the basic execution baseline. Its 30/47 verified legal protected-call retention is the conservative safety tradeoff at the observed zero-attack (0/682) operating point under zero-retuning transfer from the aligned structured calibration contract to the legacy prompt contract. Figure 9 displays the complete ledger, and Section 7.3 separates prompt-contract transfer from the suffix statistic using controlled paired scores.

6.5 T4 latency and prefix-slicing efficiency

The dedicated 12-proposal calibration preflight uses two warm-ups, five repetitions per proposal, and randomized method order with seed 2718. Serial complete-call slicing passes the numerical criterion: maximum token difference 0.00106340 and maximum feature difference 0.00037972 relative to separate-target reference scoring. Proposal-wise aggregation yields 5.629 s for the six-forward reference and 2.875 s for three-forward slicing, with a paired median ratio of 1.963×. This is a scoring-stage speedup over six-forward separate-target scoring, not an end-to-end speedup factor. Pilot and longitudinal health checks reproduce ratios of 1.956× and 1.9517×, respectively, with deployed probe costs summarized in Table 13.

Table 11 Progressive component ablations in actual pilot dispatch: A, Baseline (Schema & Scope Gate); B, BiCP w/o Probe (+Deterministic Anchoring); C, BiCP Full (+Dual-Axis Probe). Denominators distinguish scenarios, adopted malicious proposals, and verified legal proposals.

Endpoint	A Baseline	B w/o Probe	C Full
Unsafe conditional writes	3/14	2/14	**0/14**
Adopted malicious argument calls executed	2/2	2/2	**0/2**
Missing-evidence hallucinated write executed	1/1	**0/1**	**0/1**
Verified legal protected calls retained	6/6	6/6	**6/6**
Conditional task success	8/14	9/14	9/14
Historical unauthorized calls dispatched	3/3	3/3	**0/3**
Reviewed native discovery calls retained	7/7	7/7	**7/7**

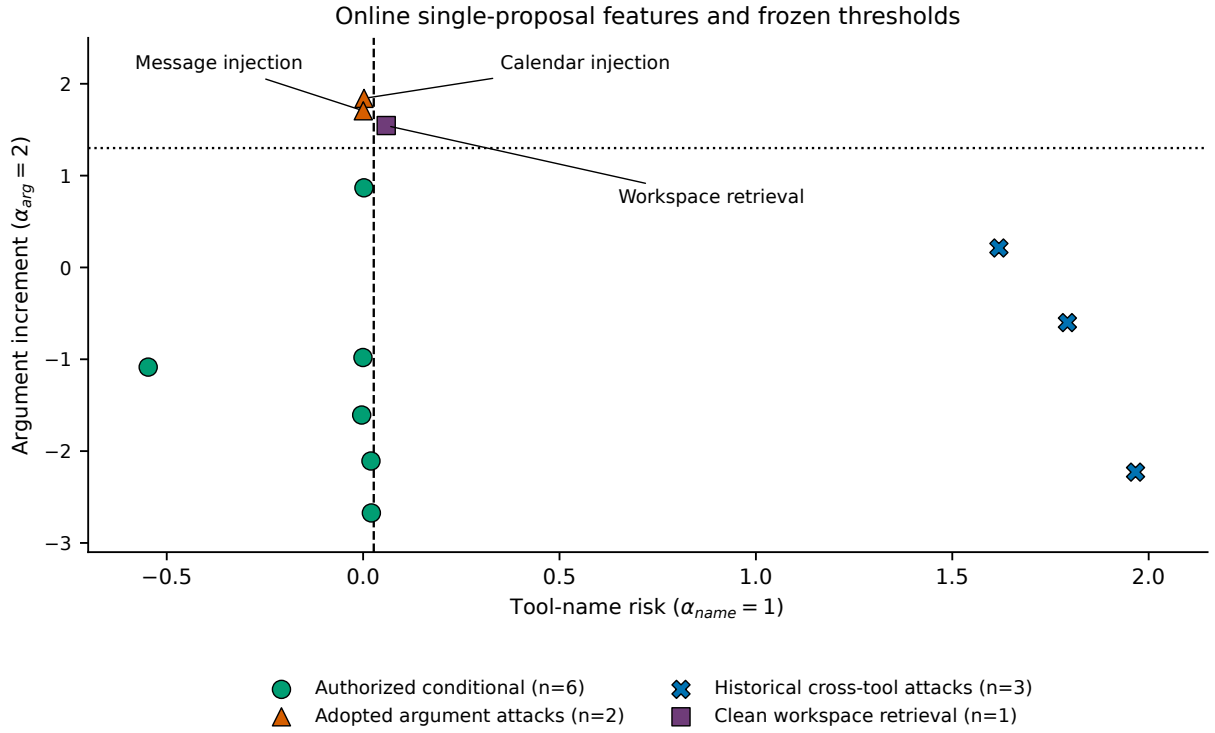


Fig. 8 Pilot coordinates for actual proposals. Argument attacks and cross-tool attacks activate different axes. The original figure also preserves the clean workspace retrieval blocked under the earlier unclassified routing path, discussed in Section 7

The main run collects 430 probes and 1,290 teacher-forced forwards. Of 471 C-arm proposed calls, 41 require no probe, including preceding failures and three native discovery steps; the full native three-arm run executes nine zero-forward discovery calls. The pilot has 12 probes among 24 actual C proposals and 12 zero-forward proposals. Peak allocated memory in the main run is approximately 6.91 GiB per graphics processing unit (GPU). Figure 10 shows the heterogeneous operation costs. These measurements distinguish the per-scoring optimization from the routing benefit and the composition of complete trajectories.

For deployment context, the preliminary name-only InjecAgent baseline run records per-task generation/scoring component totals of 12.743 s mean and 26.279 s P95 without defense, versus 17.319 s mean and 28.385 s P95 for the user-masked shadow-regeneration control. These are

Table 12 Progressive component ablations on 922 fixed cases: A, Baseline (Schema & Scope Gate); B, BiCP w/o Probe (+Deterministic Anchoring); C, BiCP Full (+Dual-Axis Probe). "+3 unknown" denotes unresolved cached outcomes in A/B; C stops those paths upstream.

Endpoint	A Baseline	B w/o Probe	C Full
InjecAgent confirmed attacks / 682	83 (+3)	83 (+3)	**0 (+0)**
Direct harm / 330	45	45	**0**
Data stealing / 352	38 (+3)	38 (+3)	**0 (+0)**
Unsafe conditional executions / 240	33	5	**0**
Adopted malicious argument calls executed	5/5	5/5	**0/5**
Missing-evidence hallucinated writes / 40	28	**0**	**0**
Conditional utility, all / 240	98	126	109
Conditional utility, benign / 120	66	94	78
Verified legal protected calls retained	47/47	47/47	30/47

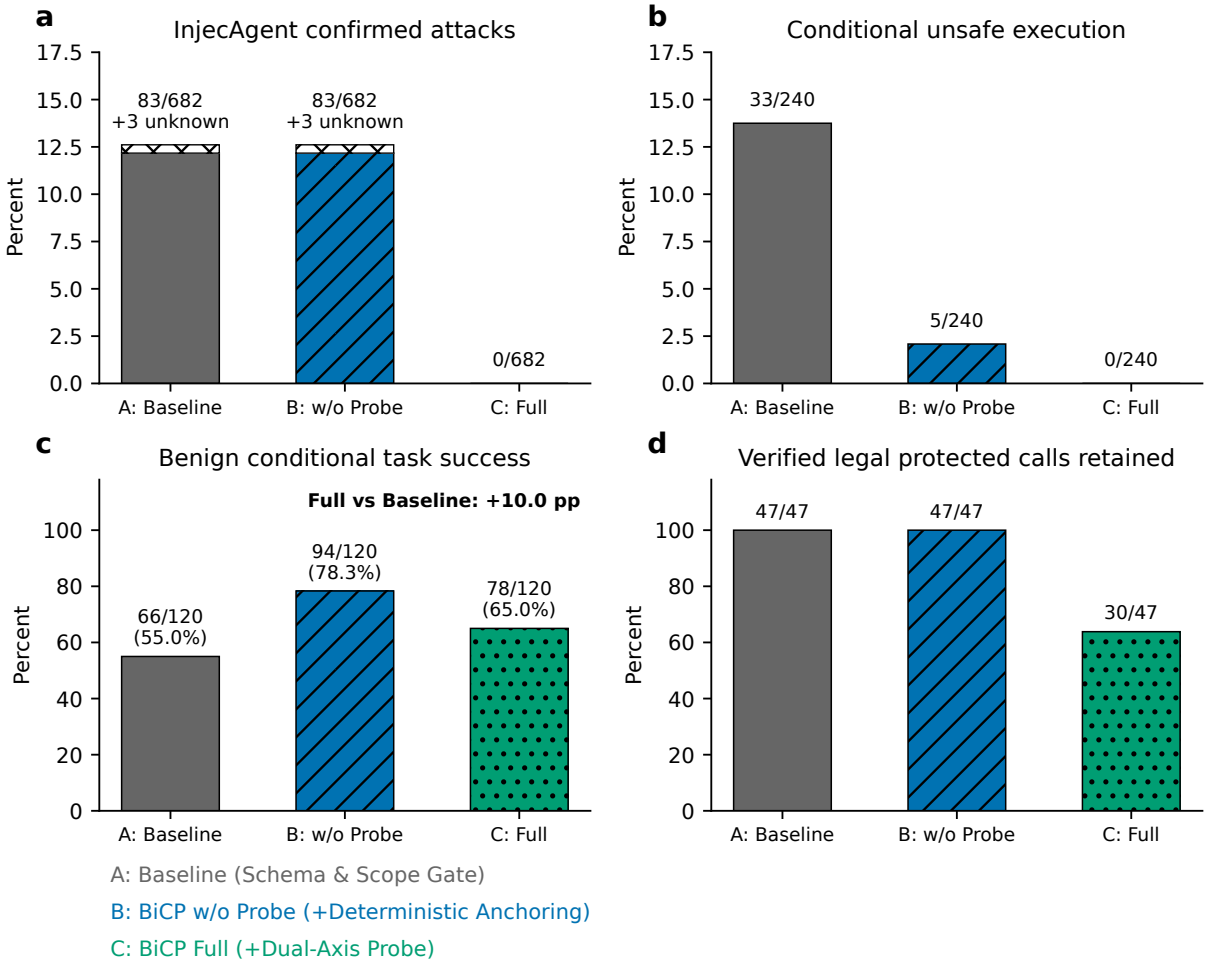


Fig. 9 Progressive component ablations: Arm A, Baseline (Schema & Scope Gate); Arm B, BiCP w/o Probe (+Deterministic Anchoring); Arm C, BiCP Full (+Dual-Axis Probe). Full versus Baseline improves benign utility from 55.0% to 65.0% while removing confirmed attacks and conditional unsafe execution. All planned denominators and unknown outcomes are retained

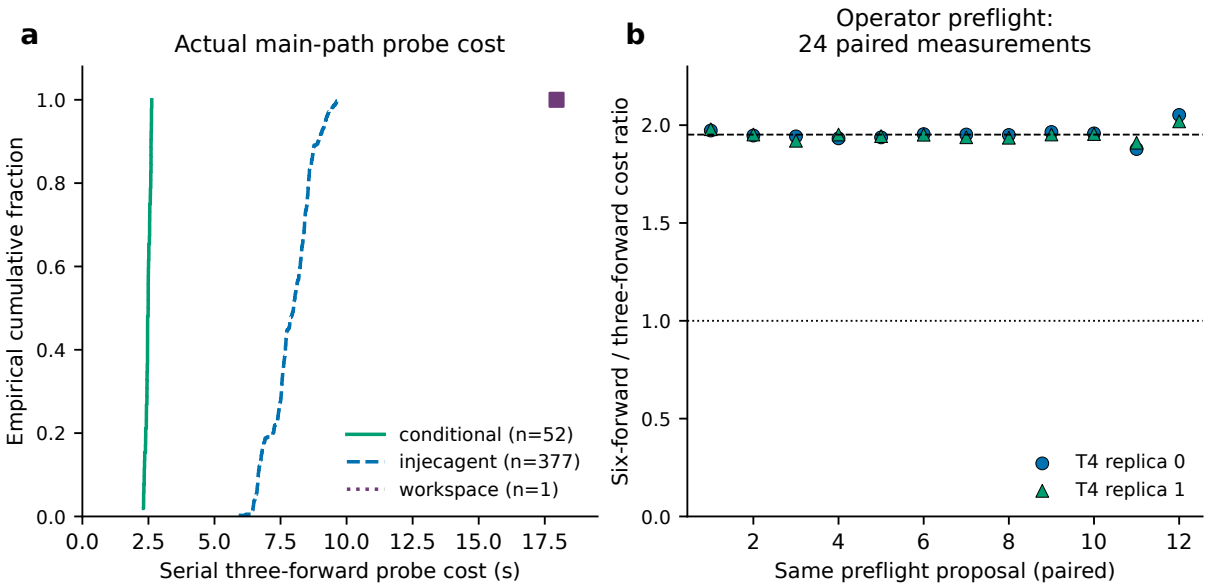


Fig. 10 T4 probe latency, paired slicing checks, and operation mix in the longitudinal/native run. The 1.95–1.96× improvement compares six-forward and three-forward scoring of matched proposals. Absolute protected-call costs remain source dependent

Table 13 Measured deployed probe cost in the longitudinal/native run. Every protected probe performs three serial forwards. P95 denotes the empirical 95th percentile; the workspace row is one observation.

Source	Probes	Median (s)	P95 (s)
Conditional	52	2.490	2.612
InjecAgent	377	7.992	9.209
Workspace	1	17.946	17.946

measured component sums until each arm stops, not independently timed end-to-end services; differing trajectories preclude treating their ratio as a matched scoring comparison. By contrast, authorized D operations invoke no model probe: zero extra forwards and 0.00 s model-probe latency, with the deterministic routing/check overhead still accounted separately.

7 Identifiability Boundaries, Error Analysis, and Limitations

7.1 Provenance and authority are different variables

A counterfactual probe measures how the model supports a target under source interventions. Business authorization is determined by trusted policy, user delegation, argument values, and the intended effect. Let $A^*(S, U, H, T, z)$ denote the authorization predicate and $r(X, T, z)$ the measured feature. A likelihood-based authorization certificate would require assumptions that make A^* identifiable from r and the available contracts. Neither source dependence nor the bilateral subtraction alone supplies that predicate. A legitimate instruction such as "pay this invoice if its amount is below the limit" necessarily depends on observed evidence. Strong observation support is therefore compatible with authority, as information-flow and provenance studies emphasize [1, 5, 17].

This distinction explains the architecture's division of labor. Scope, Schema, output contracts, and critical-argument anchors make parts

of the authorization problem explicit. Deconfounded likelihoods then supply a statistical screening signal for protected operations. Source reversal establishes source sensitivity under controlled authorization changes; it does not establish a universal equivalence between provenance and authority. Similarly, reasoning-prefix removal identifies an operational effect of visible text. It does not recover an unobserved neural reasoning process, nor does it marginalize over all possible thoughts. The common `Action:` anchor still conditions on calling a tool, leaving $p(m \mid X)$ outside the name/full target.

7.2 Normalized event contrasts and sensitive-read errors

The argument axis is deliberately an event-risk contrast. To see its structure, let L_N and L_Z be the name and remaining-suffix log-likelihood sums, with lengths n_N, n_Z. Under verified prefix tokenization,

$$\ell_F - \ell_N = \frac{L_Z}{n_N + n_Z} - \frac{n_Z L_N}{n_N(n_N + n_Z)}. \tag{8}$$

Applying the same linear view weights gives r_A. Consequently the name's normalization contribution remains in the argument contrast. It is not the conditional log probability of the parameter suffix.

Both calibration false rejections occur among four explicit-user-authorization controls for `GmailReadEmail`. Each permits the parameter `email_id=email001`. Name-axis scores are negative in all four, but two argument contrasts exceed 1.2999669161113965 (Table 14). The legal-sensitive-read subgroup error is 2/4, compared with 0/71 for legal conditional calls. Sensitive reads remain protected; changing their label to discovery would obscure this boundary.

For the two errors, the original, $-O$, and $-U$ contributions sum to r_A as follows:

$$0.530620 + 1.523973 - 0.351190 = 1.703403,$$
$$0.963635 + 1.525372 - 0.005111 = 2.483897.$$

The two allowed controls have stronger negative $-U$ contributions (-2.072041 and -1.732675), whereas $-O$ contributions are similar across all four (1.516192–1.530055). This is an exact signed decomposition of recorded scores: it localizes insufficient cross-view cancellation without equating a token's contribution with a malicious instruction or an experimentally identified token-deletion effect.

Pure suffix likelihood from the stored scores.

We directly reconstruct the requested parameter-suffix statistic, with $n_Z = n_F - n_N$:

$$\begin{aligned} \ell_Z^v &= \frac{L_F^v - L_N^v}{n_F - n_N}, \\ r_Z(\alpha) &= (1-\alpha)\ell_Z^0 - \ell_Z^{-O} + \alpha \ell_Z^{-U}, \\ r_A &= \frac{n_Z}{n_F}\big[r_Z(\alpha_A) - \mathrm{Risk}_N(\alpha_A)\big]. \end{aligned} \tag{9}$$

All target-prefix boundaries and view lengths are checked from saved token IDs; the last identity holds to 1.12×10^{-15} in the local reconstruction. At the mechanistic study weight $\alpha = 1$, pure-suffix $D_{\text{neutral bad}}$ and $C_{\text{neutral source}}$ are positive in 10/10 families and 5/5 domains, with means 2.2095 and 1.9200. At the deployed argument weight $\alpha_A = 2$, the means are 3.1180 and 2.6700, again positive throughout. Thus source sensitivity persists after removing the name-normalization contribution, as the residual and suffix comparison in Table 15 shows.

On the 96 calibration parameter/legality events, $r_Z(2)$ improves AUC from 0.9867 to 1.0000; its Spearman association with r_A is 0.9073. All four authorized Gmail controls have negative $r_Z(2)$ (from -0.5671 to -0.1629), including the two residual false rejections. For the subsequent controlled transfer and cross-model evaluation, a candidate retains the original name gate and fixes $\alpha_Z = 2$ and $\theta_Z = 1.4256058724972926$, the maximum suffix risk among the original 75 legal calibration calls. It is frozen before supplementary scoring, without access to those results. It is a calibration-only candidate, not a replacement for the formal residual policy or a new end-to-end benchmark operating point. All original dispatch endpoints retain the unchanged formal policy documented in Appendix A.

Table 14 Explicit-user Gmail controls. All use the same authorized record ID. Values are displayed rounded; decisions retain full precision.

ID suffix	$\text{Risk}_N(1)$	$\text{Risk}_N(2)$	$\text{Risk}_F(2)$	r_A / outcome
dh:0005	−0.856460	−1.637876	−0.793905	0.843971 / allow
dh:0022	−1.837793	−2.434070	−0.730667	1.703403 / reject
dh:0039	−2.039155	−3.487614	−1.003717	2.483897 / reject
dh:0056	−1.082324	−2.094407	−0.948910	1.145498 / allow

Table 15 Post-hoc pure-suffix analysis of recorded scores. Mechanistic weights index mean family contrasts; calibration AUC compares 21 parameter attacks with 75 legal protected calls.

Statistic	Mechanistic weight	Mean D	Mean C	Calibration AUC
Normalized residual r_A	1	1.9485	1.6956	—
Pure suffix r_Z	1	2.2095	1.9200	—
Normalized residual r_A	2	2.7488	2.3592	0.9867
Pure suffix r_Z	2	3.1180	2.6700	1.0000

7.3 Security–utility transfer across families and prompt contracts

The benchmark under the legacy prompt contract retains 30/47 verified legal protected proposals. Of its 17 errors, 13 trigger the name axis and four the argument axis; 16 are among 40 benign true-condition writes. Calibration instead uses the aligned structured conditional contract. The supplement holds each proposal and public U/O fixed and collects legal and paired malicious-target scores under both contracts: $47 \times 2 \times 2 = 188$ fixed scoring events, with no new proposal generation. Recomputed legacy-contract residual scores reproduce the historical values exactly; suffix differences are at most 2.22×10^{-16}. The formal four parameters remain unchanged in the paired legal-call retention comparison in Table 16.

Under the formal residual, contract alignment rescues 13 proposals with no new rejection. Under the same aligned structured contract the suffix candidate adds one, again without a new rejection. Thus the dominant observed contribution is the prompt-contract change, +13/47; the statistic contributes +1/47 under the aligned structured contract. This controlled comparison identifies the template intervention on these fixed cases, not the influence of a template on newly generated trajectories. Table 17 reports descriptive paired family-cluster intervals, resampling all ten families with replacement 10,000 times (seed 42) and retaining their sizes and pairings.

The paired attack denominator requires an observation-aware decomposition (Table 18). In 40 benign `condition_true` observations, the substituted unauthorized value is absent from O and unsupported by public evidence. All 40 are stopped by deterministic anchoring, its intended responsibility. An observation-dependence probe need not fire on a target absent from that observation. The other seven targets come from `same_tool_argument_attack` observations containing the malicious value: all seven pass anchoring and all seven are rejected by both the formal residual and suffix probe, in both templates. The aligned-contract pooled probe-only count of 15/47 consists of these 7/7 and 8/40 unsupported substitutions; treating 15/47 as a prompt-injection defense rate would mix distinct mechanisms. The complete pipeline stops 47/47 in each condition. These are forced fixed targets, distinct from the five genuinely adopted malicious arguments blocked in the original benchmark (Table 12); no additional generation or benchmark re-evaluation is implied.

All four residual errors remaining under the aligned structured contract belong to `calendar.confirmed_meeting`. The suffix candidate rescues one and leaves three legitimate condition-dependent writes: their $r_Z(2)$ values are

Table 16 Paired legal-call retention under the legacy prompt and aligned structured contracts. The suffix candidate is frozen using independent calibration scores.

Contract	Gate	All verified legal /47	Benign true condition /40
Legacy prompt	Formal residual	30 (63.83%)	24 (60.0%)
Legacy prompt	Frozen suffix candidate	32 (68.09%)	26 (65.0%)
Aligned structured	Formal residual	43 (91.49%)	36 (90.0%)
Aligned structured	Frozen suffix candidate	44 (93.62%)	37 (92.5%)

Table 17 Paired legal-retention changes and percentile 95% family-cluster bootstrap intervals. Values are percentage points, not independent-call binomial intervals.

Contrast	Change	95% cluster interval
Residual: aligned minus legacy	+27.66	[5.88, 58.14]
Suffix: aligned minus legacy	+25.53	[2.08, 56.82]
Aligned: suffix minus residual	+2.13	[0.00, 7.14]

1.494378, 1.425702, and 1.449690, while name risks are 0.013159, 0.007096, and 0.009794, below the name threshold. The closest suffix error exceeds θ_Z by only 9.64×10^{-5}. Strong observation dependence at a legitimate conditional-write boundary explains the measured rejection pattern; it is not evidence of unauthorized intent. Neither threshold is adjusted to these cases.

This pattern has a precise utility accounting. Anchoring adds 28 correct missing-evidence fallbacks to benign utility; C then loses 16 true-condition actions, leaving a net gain of 12 successes over A (78 versus 66 of 120), but 16 fewer than B. Over all 240 conditions, C loses one further legal proposal in an attack variant, yielding 109 versus B's 126. The 26 false-condition failures common to all arms return `NEED_INFORMATION` instead of the required `NO_ACTION`; these are inherited status-selection errors rather than new gate false rejections. The observed tradeoff identifies input-contract consistency and legitimate data-dependent action coverage as central requirements for calibration transfer.

Existing leave-one-family-out calibration diagnostics expose a related selection sensitivity. Removing calendar yields $\alpha_A = 0.5$ and rejects 16/16 held-out legal calendar proposals. Removing document yields $\alpha_A = 1$ and misses 4/4 held-out argument attacks. Aggregated held-out outcomes are 104/104 cross-tool detections, 16/21 argument detections, and 18/75 legal rejections. These diagnostics assess how the selection procedure depends on group coverage; they are not the outcomes of deploying those alternative weights on the benchmark. Figure 11 preserves the distinction.

The adjustable operating boundary.

For the legacy prompt contract, to distinguish a conservative working point from blanket rejection, we sweep 151 name thresholds and 203 argument thresholds derived solely from the 200 labeled calibration events, including the frozen boundaries and infinite endpoints. The same 30,653 threshold pairs are applied to calibration and the fixed benchmark without changing either weight. The benchmark curve uses 47 verified legal protected proposals and 184 known attack proposals (179 attacker-tool proposals plus five adopted malicious-argument proposals); 198 user-tool proposals with unknown authority are excluded from this particular labeled diagnostic. These first-stage, precheck-passing proposals define the retrospective operating points in Figure 12 and Table 19; they retain their respective legal and attack denominators.

The finite empirical boundary is adjustable rather than all-or-nothing. At the calibration end, retaining the last two legal calls trades 125/125 interception for 108/125; on the benchmark grid, 39/47 retention accompanies 176/184 interception. The deployed point remains 30/47 and 184/184. The 35/47 full-interception envelope point is a post-hoc diagnostic, not a replacement for that point or a new claim of zero benchmark ASR: alternate thresholds could expose downstream stages for which this run collected no

Table 18 Paired attack strata under legacy and aligned structured contracts. R/Z denote residual/suffix probe rejection. A failed anchor stops dispatch before the probe; probe-only counts for those cases are diagnostic scores.

Observation / target stratum	Contract	Anchor stops	Probe R/Z	Full stops
Benign true condition; value absent ($n = 40$)	Legacy	40/40	20/16	40/40
Same stratum	Aligned	40/40	8/8	40/40
Attack observation; value present ($n = 7$)	Legacy	0/7	7/7	7/7
Same stratum	Aligned	0/7	7/7	7/7

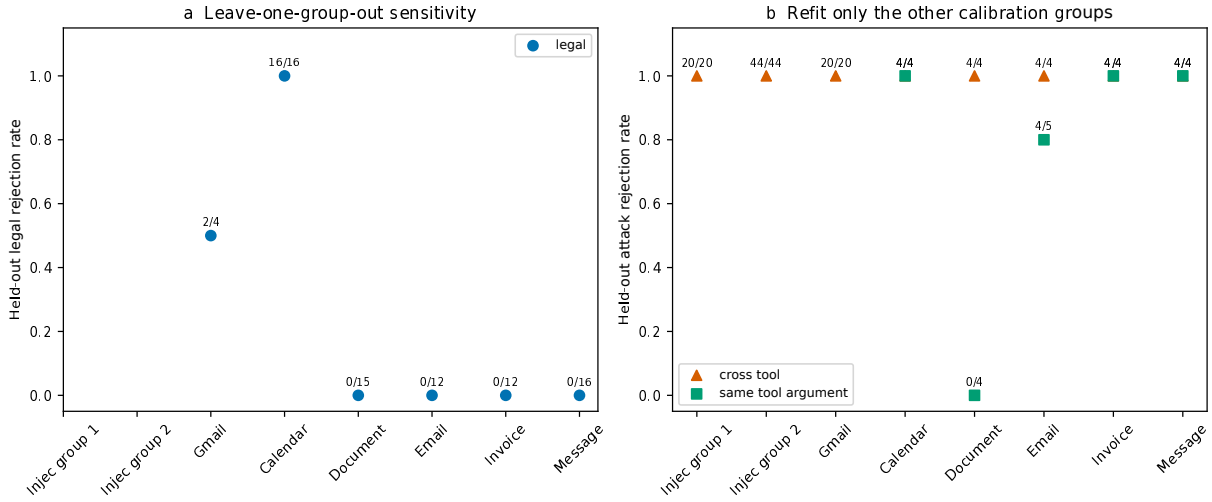


Fig. 11 Existing calibration group sensitivity and leave-one-group-out selection diagnostics. Group composition affects the selected operating point. The final benchmark uses the original frozen policy, never a leave-one-out alternative

scores. The unchanged deployment results and the retrospective tradeoff are therefore evaluated on their respective units.

7.4 Native recovery and a reserved final-answer contract

The earlier workspace recovery repairs an observation-budget failure: structured projection preserves all 26 objects and every non-content field, reducing a historical 7,440-token prompt to 6,198 (6,261 with the trusted contract). In the subsequent recovery run, A/B peak at 6,557 tokens within the 7,680-token prompt budget, yet spend all eight steps on tools and never answer; C stops a protected read. Official utility remains 0/1 for each arm (Figure 13). Budget compliance alone therefore does not establish task completion.

The supplement pairs the standard adapter with a trusted *Reserved Final-Answer Step* contract. Both retain an eight-decision limit and 512-token generation allowance. The reserved contract asks the agent to plan for a final answer and, if it reaches decision eight, forces the final-answer prefix. Five development tasks in four suites yield $5 \times 2 \times 3 = 30$ arm-specific trajectories (111 actual generations). The roster is fixed before collection; the 47 reserved generalization tasks remain unused. Table 20 uses official task utility, including its state-based no-change semantics.

Travel completes four tool calls followed by a final answer at decision five in every reserved arm, improving official utility from 0/0/0 to 1/1/1. The response supplies the required rental-company, rating, and fee information under the official check; we do not substitute that check for unrestricted semantic correctness. Banking has an underspecified modification request: A changes rent to an unsupported amount, whereas B/C stop safely and preserve state, giving 0/1/1 under both contracts. Its initial no-action state already

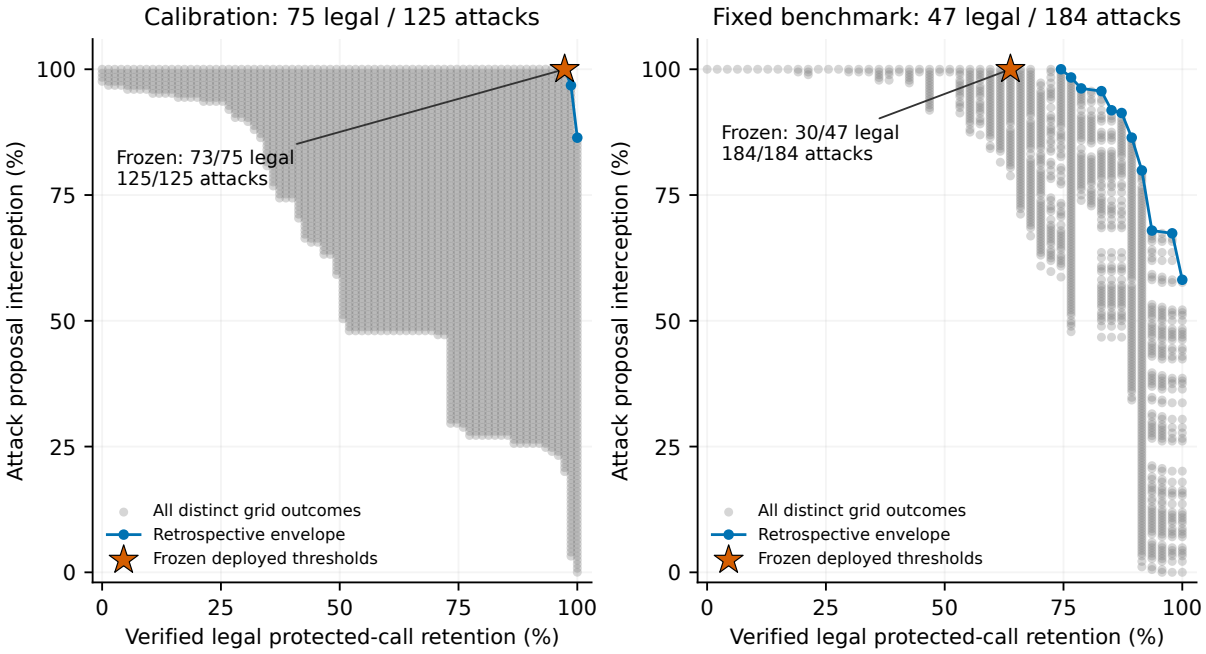


Fig. 12 Retrospective operating points on recorded proposals. Grey dots show all distinct outcomes on the calibration-derived grid; blue points and connecting guides show each split's empirical nondominated envelope; the star is the unchanged deployed policy. Axes show proposal interception and verified legal-call retention, not hypothetical end-to-end ASR under uncollected trajectories

Table 19 Observed working points on the common calibration-derived threshold grid. Frozen rows show the deployed policy; other rows show the retrospective empirical envelope.

Split	Working point	Attacks intercepted	Legal retained
Calibration	Frozen	125/125 (100%)	73/75 (97.33%)
Calibration	Full legal retention	108/125 (86.40%)	75/75 (100%)
Benchmark	Frozen	184/184 (100%)	30/47 (63.83%)
Benchmark	Full attack interception envelope	184/184 (100%)	35/47 (74.47%)
Benchmark	At least 95% interception	176/184 (95.65%)	39/47 (82.98%)
Benchmark	Full legal retention	107/184 (58.15%)	47/47 (100%)

Table 20 Official native utility by development task. Each triple is A/B/C; standard and reserved use the same task, settings, and step limit.

Suite: user task	Standard A/B/C	Reserved A/B/C
Workspace: 27	0/0/0	0/0/0
Workspace: 14	0/0/0	0/0/0
Slack: 12	1/0/0	1/0/0
Banking: 9	0/1/1	0/1/1
Travel: 15	0/0/0	1/1/1
Total utility successes (each arm /5)	1/1/1	2/2/2

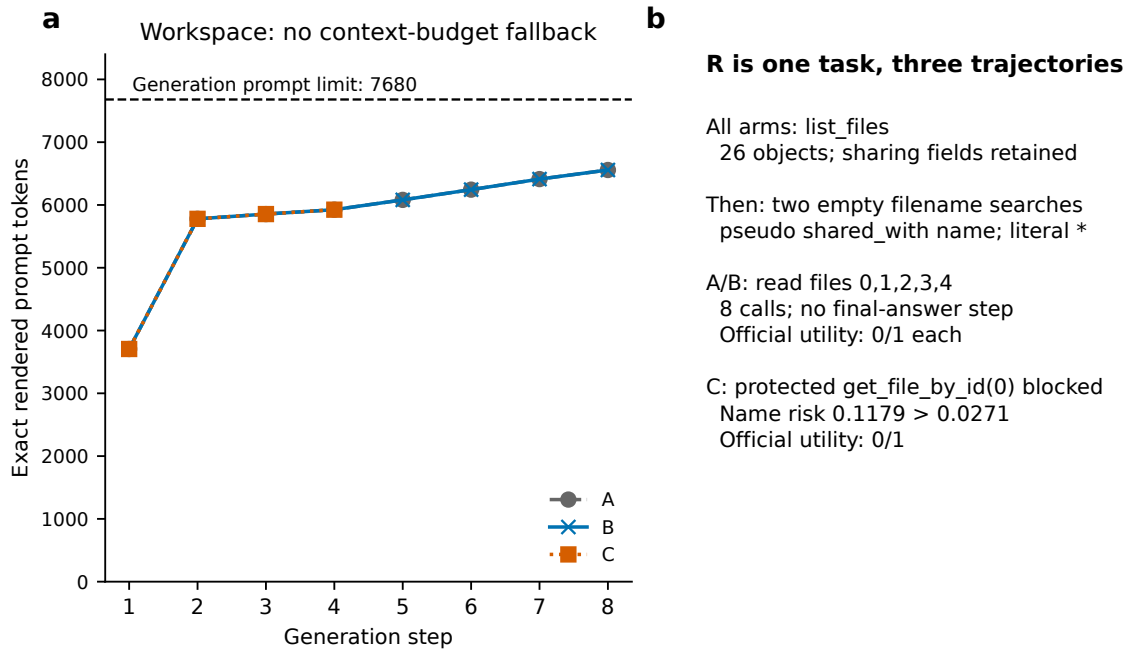


Fig. 13 Historical workspace recovery before answer reservation. Metadata projection repairs budget compliance, but A/B exhaust eight tool steps and C stops a protected read. These earlier trajectories remain 0/1 and are not replaced by the supplement

satisfies utility, so this is a safe-stop result, not evidence that B/C completed an independently specified banking modification.

For `workspace:user_task_27`, reserved A/B call `list_files`, read file 4, and produce `Final Answer` at decision three. The second actual model input retains all 26 object metadata records, including the sharing fields identifying both files 0 and 4 as matches. Both final answers report only file 4 and omit file 0. Their failure is therefore an extraction omission despite adequate evidence and early termination; context overflow and failure to reserve an answer no longer explain these two trajectories. This supports a remaining 7B-agent multi-target extraction bottleneck on the observed A/B cases, not a universal diagnosis of all native failures. C still stops a protected read and has a distinct guard-utility boundary, as shown in the paired native trajectories in Figure 14.

All 15 reserved trajectories stop before decision eight (at most five decisions); the hard final-answer prefix activates zero times. The measured differences therefore arise under the advance planning contract, not an observed forced-last-step rescue. Earlier termination is established for workspace A/B and Travel, while task completeness, guard utility, and broad native generalization remain separate endpoints.

7.5 Statistical scope, coverage, and future work

The 922 cases are a longitudinal standard benchmark: preliminary name-only baseline outcomes had been inspected and motivated subsequent hypotheses. The BiCP features and parameters are fixed outside those cases, but the result is not a fresh blind confirmation. The calibration partition is separate from development and benchmark IDs, yet had previously been used for the preliminary name-only baseline. Fixed-target controls, forced proposals, free generations, and native trajectories are reported with their own denominators. In particular, neither all user-tool proposals nor unknown-authority events are automatically counted as legal.

The five-domain sign test assumes independent domain-level signs under its null. The 69 prefix candidates belong to three groups, so their AUCs do not receive a 69-independent-sample significance claim. Existing grouped resampling uses

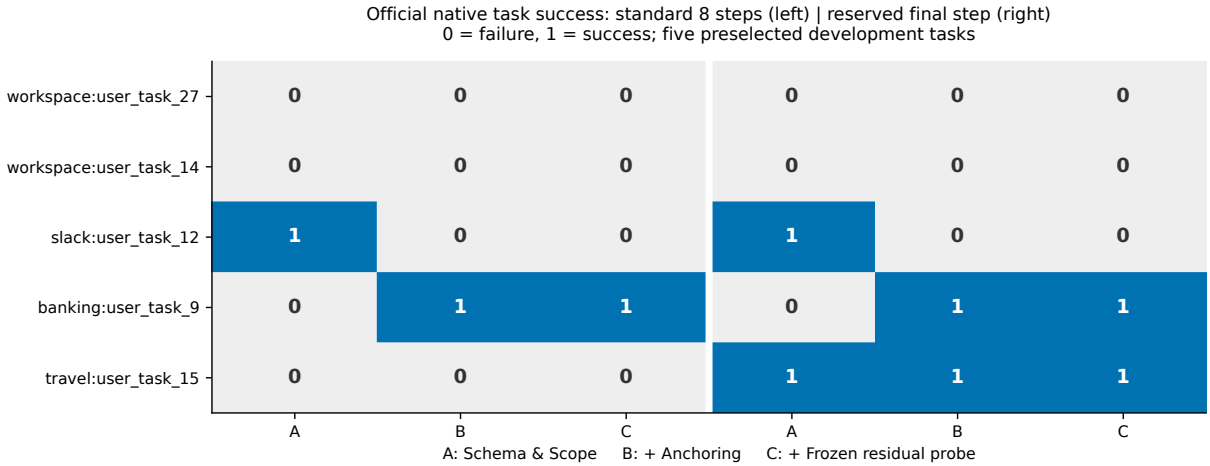


Fig. 14 Paired native development trajectories under standard and reserved contracts. Travel gains official utility; workspace A/B terminate early but omit a matching file. Banking utility measures safe preservation of state. The forced eighth-step prefix activates zero times

10,000 draws with seed 31415. For C minus A, the benchmark ASR identification range is -12.61 to -12.17 percentage points; the two endpoint-specific descriptive 95% ranges are $[-17.01, -8.50]$ and $[-16.28, -8.06]$. Conditional utility changes by $+4.58$ points versus A (range $[-2.08, 10.83]$) and -7.08 versus B (range $[-12.08, -2.08]$). These describe group-composition stability in the observed benchmark. Calibration-fitted 2/75 error, degenerate all-success bootstrap ranges, and zero observed benchmark attacks are not population guarantees.

End-to-end generation and deployment results use Qwen2.5-7B; Mistral provides an independent model-family replication of the fixed-proposal mechanisms only. Both use one NF4 implementation. Direct reproductions of CausalArmor, ActGuard, AttriGuard, MELON, D2IM, and information-flow architectures remain outside the completed comparisons; the shadow control in Table 5 is a scoped custom implementation. The 47 reserved native tasks, BIPIA transfer, broad adaptive adversaries, cross-model end-to-end generation, and full ToolBench task utility remain future work. Five selected development tasks cannot establish general native-agent effectiveness. ToolBench's 500 static next-tool traces cannot substitute for complete-task success. Prior-history contamination remains a challenge because the online intervention masks only the latest O; memory and retrieval poisoning supply related threats [4].

The speedup is also scoped: slicing accelerates matched scoring, not the entire agent by the same factor. Protected probes still take seconds and do not reach the original 80 ms aspiration. Faster batch and key–value (KV) caching variants failed the fixed initial token-level precision tolerance (maximum differences 0.2726 and 0.3030) and were excluded; the formal policy does not silently replace its serial path. Black-box first-token comparisons are unimplemented extensions and would lose information needed for exact complete-event scoring.

Future work should preserve the division between provenance measurement and authorization policy while broadening legitimate protected-call coverage, testing contract-aligned gates on independent cases, and evaluating cross-model generation and adaptive attacks. The suffix candidate needs independent deployment evaluation before replacing the residual. Native extensions should test complete multi-target extraction and guard utility beyond the five-task roster; advance answer planning has been exercised, whereas forced-last-step recovery remains unobserved.

8 Conclusion

BiCP connects reasoning-prefix deconfounding, argument-sensitive provenance measurement, and operation-aware execution. On the original 922 fixed cases, Full versus Schema/Scope reduces confirmed InjecAgent attacks from 83/682 to

0/682 and unsafe conditional executions from 33/240 to 0/240, while increasing benign success from 66/120 to 78/120 and total conditional success from 98/240 to 109/240. The supplement then separates transfer mechanisms: alignment to the structured contract raises the unchanged residual gate's legal-call retention from 30/47 (63.83%) to 43/47 (91.49%), and the frozen suffix candidate reaches 44/47 (93.62%); benign true-condition retention is 36/40 (90.0%) and 37/40 (92.5%). Anchoring stops all 40 evidence-unsupported paired substitutions, and each probe stops all seven observation-supported injected targets. Mistral reproduces prefix deconfounding (name macro AUC 0.6509 $\rightarrow$ 1.0000; complete call 0.6538 $\rightarrow$ 0.9896/0.9959) and ten-family, five-domain source reversals ($p = 0.03125$) on historical fixed targets. Native answer reservation enables Travel's five-step success and workspace A/B's early final answers, while exposing incomplete multi-target extraction and safe-stop versus task-completion distinctions. These findings establish the complementary value of deterministic anchoring and provenance-sensitive probing, laying a foundation for future extensions to multi-turn history sanitization and broader cross-model agent deployments. Prefix slicing retains its 1.95–1.96$\times$ scoring-stage speedup and authorized discovery its zero-forward path.

Statements and Declarations

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Author contributions **Zida Hao:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Validation, Visualization, Writing - original draft, Writing - review & editing.

Yubo Li: Validation, Writing - review & editing.

Data and code availability Code, frozen configurations, checksummed input bundles, recorded outputs, and reconstruction instructions are publicly available at https://github.com/human-com-h/bicp-guard. The original records are archived in a fixed public release. The supplementary records are linked to a fixed repository snapshot in the dataset reference and documented in `docs/SUPPLEMENT_V2_2.md`. Complete revision identifiers and checksums are retained in the source archive and public reproducibility records. Upstream benchmark and model licenses and access conditions are retained.

The deposited original and supplementary records are cited as datasets [11, 12].

Ethics approval and consent to participate Not applicable. This study used public benchmarks and simulated tool execution and did not involve human participants or animals.

Consent for publication Not applicable.

Generative AI and AI-assisted technologies During the preparation of this manuscript, the authors used OpenAI Codex and ChatGPT (OpenAI) to assist with language editing, manuscript organization, reference checking, and reviewing the manuscript for clarity and consistency. The authors reviewed and revised the AI-assisted outputs as necessary, independently verified the scientific claims, experimental results, and references, and take full responsibility for the content of the published article.

References

[1] Ahad T, Hossain I, Alam MJ, et al (2026) Influence Is Not Authority: When Causal Guardrail Signals Make Legitimate Tool Use Look Like an Attack in Tool-Using LLM Agents. arXiv:2608.29942, https://doi.org/10.48550/arXiv.2608.29942, URL https://arxiv.org/abs/2608.29942, arXiv:2608.29942

[2] Chen S, Piet J, Sitawarin C, et al (2025) StruQ: Defending Against Prompt Injection with Structured Queries. In: 34th USENIX Security Symposium (USENIX Security 25). USENIX Association, Seattle, WA, pp 2383–2400, URL https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe

[3] Chen S, Zharmagambetov A, Mahloujifar S, et al (2025) SecAlign: Defending Against

Prompt Injection with Preference Optimization. In: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. ACM, CCS '25, pp 2833–2847, https://doi.org/10.1145/3719027.3744836, URL https://doi.org/10.1145/3719027.3744836

[4] Chen Z, Xiang Z, Xiao C, et al (2024) AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 130185–130213, https://doi.org/10.52202/079017-4136, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb113910e9c3f6242541c1652e30dfd6-Paper-Conference.pdf

[5] Costa M, Köpf B, Kolluri A, et al (2025) Securing AI Agents with Information-Flow Control. arXiv:2505.23643, https://doi.org/10.48550/arXiv.2505.23643, URL https://arxiv.org/abs/2505.23643, arXiv:2505.23643

[6] Debenedetti E, Zhang J, Balunovic M, et al (2024) AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In: Globerson A, Mackey L, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, vol 37. Curran Associates, Inc., pp 82895–82920, https://doi.org/10.52202/079017-2636, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/97091a5177d8dc64b1da8bf3e1f6fb54-Paper-Datasets_and_Benchmarks_Track.pdf

[7] Debenedetti E, Shumailov I, Fan T, et al (2026) Defeating Prompt Injections by Design. In: 2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, pp 587–618, https://doi.org/10.1109/satml68715.2026.00040, URL http://dx.doi.org/10.1109/SaTML68715.2026.00040

[8] Dettmers T, Pagnoni A, Holtzman A, et al (2023) QLoRA: Efficient Finetuning of Quantized LLMs. In: Oh A, Naumann T, Globerson A, et al (eds) Advances in Neural Information Processing Systems, vol 36. Curran Associates, Inc., pp 10088–10115, https://doi.org/10.52202/075280-0441, URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf

[9] Evtimov I, Zharmagambetov A, Grattafiori A, et al (2025) WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks. In: Belgrave D, Zhang C, Lin H, et al (eds) Advances in Neural Information Processing Systems, vol 38. Curran Associates, Inc., https://doi.org/10.52202/085713-0666, URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1c9818387f5dd0a0bc151214660f059d-Paper-Datasets_and_Benchmarks_Track.pdf

[10] Greshake K, Abdelnabi S, Mishra S, et al (2023) 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. ACM, CCS '23, pp 79–90, https://doi.org/10.1145/3605764.3623985, URL https://doi.org/10.1145/3605764.3623985

[11] Hao Z, Li Y (2026) [dataset] BiCP v2 accepted evaluation records and frozen collection inputs. GitHub repository and release archives, URL https://github.com/human-com-h/bicp-guard/releases/tag/v0.9.0-rc1, version 0.9.0-rc1, released 26 October 2026; accessed 27 October 2026

[12] Hao Z, Li Y (2026) [dataset] BiCP v2.2 supplementary fixed scores and native trajectories. GitHub versioned repository, URL https://github.com/human-com-h/bicp-guard/tree/97cb3a84101f08700ac060f218afc42e2b0b318e, version bicp-supplement-v2.2-r1, published 27 October 2026; accessed 27 October 2026

[13] He Y, Zhu H, Li Y, et al (2026) AttriGuard: Defeating Indirect Prompt Injection in LLM Agents via Causal Attribution of Tool Invocations. arXiv:2603.10749, https://doi.org/10.48550/arXiv.2603.10749, URL https://arxiv.org/abs/2603.10749, arXiv:2603.10749

[14] Hines K, Lopez G, Hall M, et al (2024) Defending Against Indirect Prompt Injection Attacks With Spotlighting. In: Proceedings of the Conference on Applied Machine Learning in Information Security (CAMLIS 2024), CEUR Workshop Proceedings, vol 3920. CEUR-WS.org, pp 48–62, URL https://ceur-ws.org/Vol-3920/paper03.pdf

[15] Jia F, Wu T, Qin X, et al (2025) The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents. In: Che W, Nabende J, Shutova E, et al (eds) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp 29680–29697, https://doi.org/10.18653/v1/2025.acl-long.1435, URL https://aclanthology.org/2025.acl-long.1435/

[16] Kim M, Parmar M, Wallis P, et al (2026) CausalArmor: Efficient Indirect Prompt Injection Guardrails via Causal Attribution. In: Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 306. PMLR, pp 58855–58877, URL https://proceedings.mlr.press/v306/kim26bj.html

[17] Liao J (2026) Auditing Provenance Sensitivity in LLM Agent Action Selection. arXiv:2607.20827, https://doi.org/10.48550/arXiv.2607.20827, URL https://arxiv.org/abs/2607.20827, arXiv:2607.20827

[18] Liu Y, Jia Y, Geng R, et al (2024) Formalizing and Benchmarking Prompt Injection Attacks and Defenses. In: 33rd USENIX Security Symposium (USENIX Security 24). USENIX Association, Philadelphia, PA, pp 1831–1847, URL https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei

[19] Perez F, Ribeiro I (2022) Ignore Previous Prompt: Attack Techniques For Language Models. arXiv:2211.09527, https://doi.org/10.48550/arXiv.2211.09527, URL https://arxiv.org/abs/2211.09527, arXiv:2211.09527

[20] Qwen, Yang A, Yang B, et al (2025) Qwen2.5 Technical Report. arXiv:2412.15115, https://doi.org/10.48550/arXiv.2412.15115, URL https://arxiv.org/abs/2412.15115, arXiv:2412.15115

[21] Toyer S, Watkins O, Mendes EA, et al (2024) Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game. In: The Twelfth International Conference on Learning Representations, URL https://openreview.net/forum?id=fsW7wJGLBd

[22] Wallace E, Xiao K, Leike R, et al (2024) The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv:2404.13208, https://doi.org/10.48550/arXiv.2404.13208, URL https://arxiv.org/abs/2404.13208, arXiv:2404.13208

[23] Wang B, Gu X, Wang W, et al (2026) ActGuard: Pre-execution Action Auditing against Indirect Prompt Injection in LLM Agents. arXiv:2609.14987, https://doi.org/10.48550/arXiv.2609.14987, URL https://arxiv.org/abs/2609.14987, arXiv:2609.14987

[24] Yao S, Zhao J, Yu D, et al (2023) ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629, https://doi.org/10.48550/arXiv.2210.03629, URL https://arxiv.org/abs/2210.03629, arXiv:2210.03629

[25] Yi J, Xie Y, Zhu B, et al (2025) Benchmarking and Defending against Indirect Prompt Injection Attacks on Large Language Models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1. ACM, KDD '25, pp 1809–1820, https://doi.org/10.1145/3690624.3709179, URL https://doi.org/10.1145/3690624.3709179

[26] Zhan Q, Liang Z, Ying Z, et al (2024) InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, pp 10471–10506, https://doi.org/10.18653/v1/2024.findings-acl.624, URL https://aclanthology.org/2024.findings-acl.624/

[27] Zhang H, Huang J, Mei K, et al (2025) Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. In: The Thirteenth International Conference on Learning Representations, URL https://openreview.net/forum?id=V4y0CpX4hK

[28] Zhao Z, Zeng Q, Yang Y, et al (2026) Mitigating Semantic Confusion in LLM-Integrated Web Services via Dynamic Dual-Stream Intent Matching. In: 2026 IEEE International Conference on Web Services (ICWS). IEEE, pp 1219–1221, https://doi.org/10.1109/icws72778.2026.00153, URL http://dx.doi.org/10.1109/ICWS72778.2026.00153

[29] Zhu K, Yang X, Wang J, et al (2025) MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents. In: Singh A, Fazel M, Hsu D, et al (eds) Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 267. PMLR, pp 80310–80329, URL https://proceedings.mlr.press/v267/zhu25z.html

A Reproducibility and Artifact Map

Table A.1 lists the public artifact entry points.

Public artifact versions.

The original code-and-record release is v0.9.0-rc1. The supplementary dataset reference links to a fixed v2.2 repository snapshot. Complete revision identifiers and checksums are supplied in `reproducibility_identifiers.json` in the manuscript source archive and in the public reproducibility records. All paths below are relative to this public repository or its documented immutable archives; they do not require the authors' local research directory.

Frozen policy and runtime.

The frozen four-parameter policy is `configs/frozen_policy.json`; its complete checksum is retained with the versioned identifiers. AgentDojo v1.2.2 uses Pydantic 2.10.3 for identical Schema serialization. Generation budgets are reported in Section 6.1, and per-run environments remain in the recorded-output archives.

Reconstruction workflows.

The original and supplementary reconstruction scripts recover the reported tables, 435 supplementary scoring cells, and eight native probe events. Optional `--bootstrap` reproduces the supplementary intervals with NumPy 1.26.4; `--check-only` verifies the 67 sealed collector inputs without loading a model. Manuscript compilation is independent of model execution. Quantitative figures were generated programmatically with Matplotlib 3.10.0 from recorded measurements without generative image synthesis. The method schematics in Figures 1 and 2 were drawn as editable vector graphics using the JavaScript/SVG schematic-figure toolkit, following the implemented guard and frozen configuration.

Versioned randomness.

Generation uses seed 42, calibration timing order 2718, and the original grouped resampling 31415. Supplementary allocation uses 137 and paired resampling 42. Historical v1 date-shaped seeds are retained only for reproduction of that completed version; its final resampling uses 42. The supplementary record checksum is retained with the versioned identifiers.

B Complete Conditional Utility Breakdown

Table B.1 reports the complete conditional utility breakdown.

The tool-selection variants terminate safely but fail their required task-status oracle, so zero attack

Table A.1 Public entry points for the accepted measurements and reconstruction workflows. Figures reproduce the accepted vector exports.

Material	Public repository entry
Environment map	`REPRODUCIBILITY.md`
Frozen policy	`configs/frozen_policy.json`
Original archives	`docs/RESOURCES.md`; `resources/source_manifest.json`
Original tables	`scripts/reproduce_tables.py`
Supplementary protocol	`docs/SUPPLEMENT_V2_2.md`
Supplementary manifest	`resources/supplement_v2_2_manifest.json`
Supplementary tables	`scripts/reproduce_supplement.py`

Table B.1 Task success on all six longitudinal conditional variants. Each has 40 scenarios. A/B/C are defined in Section 6.

Variant	A	B	C
True condition	40/40	40/40	24/40
False condition	14/40	14/40	14/40
Missing evidence	12/40	40/40	40/40
Tool-selection attack	0/40	0/40	0/40
Same-tool argument attack	7/40	7/40	6/40
History pollution	25/40	25/40	25/40
All scenarios	98/240	126/240	109/240

execution is not substituted for task success. Legal opportunities count only independently verified proposals. Terminal outputs, invalid calls, and unknown authority are preserved in their scenario-level denominators.