On a fixed 7-8B open model and a standard indirect-prompt-injection tool-use ben

Humanity First Research

Abstract

Tool-using language-model agents read text they must not obey: an imperative planted in a web page, an email, or a tool result can hijack the episode, and this indirect prompt injection remains the central unsolved security problem for agentic deployment. Existing defenses impose the instruction-data distinction from outside, by marking inputs, finetuning a privilege hierarchy, or containing the model behind information-flow policy. We asked whether a stock instruction-tuned model already encodes that distinction internally: when the same imperative sits in the trusted prompt versus inside a tool return, is the difference realized as a low-rank residual-stream subspace through which injection-following is routed? We preregistered a test on Llama-3.1-8B-Instruct: an instruction-ness subspace of rank at most five, learned at a locked mid-layer band (layers 12 through 18) from matched contrast pairs that vary only the provenance of an imperative, projected out solely at tool-content token positions, with rank and coefficient frozen after selection on a held-out validation split.

The hypothesis is rejected, and the rejection is now bounded rather than merely unsupported. The manipulation check failed outright: a linear probe reading trusted-versus-untrusted provenance at the intervened positions scored AUC 1.0 on held-out pairs both before and after projection, a decrease of 0.0 against a required 0.2, so the intervention never removed the signal it targets. At the frozen operating point the headline comparison, rerun at n=250, moved attack success from 0.428 to 0.436; the paired bootstrap difference, defined as baseline minus intervention, is -0.008 with a 95 percent interval of [-0.056, 0.040] at $p = 0.814$, placing the largest reduction consistent with the data at 4.0 percentage points against a preregistered requirement of at least 20. Benign task cost at n=150 was exactly zero.

Two further results explain why. Sweeping the projection rank from 1 to 150 never collapsed the provenance probe: AUC remained 1.0 at every rank while up to 73.5 percent of the activation norm was removed, so provenance is not a low-rank removable direction but a redundantly encoded one. That finding survives the obvious objection. A diagnostic showed our original contrast pairs were positionally confounded, with trusted and untrusted spans occupying disjoint token-position ranges and a position-and-length-only probe reaching AUC 1.0 on its own, so we rebuilt the pairs matched by construction. On the matched set the position-only probe falls to 0.5097, at chance, while the activation probe still reads provenance at AUC 1.0 and still does not collapse under projection. Provenance is therefore genuinely and redundantly encoded, not an artifact of how the stimuli were laid out, though perfect separability under position control still does not establish that the probe reads semantic trust rather than channel and role formatting inherent to a rendered tool response.

Three methodological findings travel beyond this study. Selecting an ablation coefficient by attack-success reduction alone chose a setting that drove

Preprint. Under review.

attack success to exactly zero at every rank by destroying output coherence, so a fluency constraint is not optional. The mechanistic and behavioral measures dissociate: the probe cannot distinguish the fitted subspace from a norm-matched random one at any rank, yet at rank 20, well outside the preregistered range and reported as exploratory, the fitted subspace does lower attack success relative to random by a margin whose interval excludes zero. And our preregistered adaptive attacker could not be evaluated at all: a safety-trained assistant refused 37 of 40 escalation requests, and because the harness accepted any non-empty output as a rewrite, those refusals were fed back as injections and drove attack success toward zero in both the defended and undefended arms. Refusal detection belongs in any red-team harness that uses an aligned model as the attacker, or refusal will masquerade as defensive success.

Limits worth stating plainly: one model family, one benchmark rendered as single-turn tool-conditioned episodes, control arms and the rank sweep still at n=40 with the seed check covering two of eight preregistered ranks, and a comprehension control that lost 5.0 points against its own 5-point bound. What we claim is narrow and, we think, useful: this defense does not work at the scale tested, the reason is that the signal it targets is redundant rather than low-rank, and several of the traps we fell into are ones the next study can avoid by construction.

Hypothesis

On a fixed 7-8B open model and a standard indirect-prompt-injection tool-use benchmark with pre-checked baseline ASR >=40% and baseline benign success >=60% (scored by a calibrated injection judge, not substring matching), projecting out a k<=5-dimensional residual-stream 'instruction-ness' subspace - learned at locked mid-layers via directional projection removal, from matched contrast pairs that hold tool, topic, length, formatting, and propositional content fixed while varying only whether an imperative sits in trusted prompt tokens versus untrusted tool/data tokens, with the subspace, layer band, and coefficient all fit on a calibration/validation split disjoint from every eval injection phrasing, and applied ONLY at retrieved/tool-content token positions - reduces held-out cross-attack-family injection ASR by >=50% relative AND >=20 absolute percentage points (paired-bootstrap CI separated from zero), beats a norm-matched random-subspace ablation at the same positions and same k by a CI-separated margin, costs <=5 points on ordinary benign tasks AND on a comprehension-preserving control set whose data spans contain imperative text that must be understood but not obeyed, retains >=30% relative ASR reduction against a preregistered adaptive paraphrase attacker optimizing against the frozen intervention, and reaches >=90% of its k=20 maximum ASR reduction by k<=5 in at least 2 of 3 seeds - evidencing a genuinely low-rank, position-addressable injection-following channel rather than distributed compliance, generic activation damage, or a data-reading lesion.

Introduction

Language-model agents increasingly act through tools: they read email, browse the web, query internal systems, and take actions on a user's behalf (Arunkumar et al. 2026). Every tool call that returns attacker-influenceable content opens the same door. The model reads text it was never meant to obey, and an imperative planted in that text can hijack the episode. This is indirect prompt injection (Greshake et al. 2023), the tool-mediated descendant of direct goal hijacking (Perez and Ribeiro 2022), now formalized and benchmarked as an attack class in its own right (Liu et al. 2023) and widely regarded as the central unsolved security problem for agentic deployment (Dehghantanha and Homayoun 2026). Standard benchmarks measure it at scale: AgentDojo scores injected-task execution jointly with benign utility inside live agent loops (Debenedetti et al. 2024), and InjecAgent catalogs injection cases across tool-integrated agents (Zhan et al. 2024).

Existing defenses approach the problem from three directions. Input-side methods make provenance visible in the prompt itself, by spotlighting-style marking of untrusted spans (Hines et al. 2024) or by structured query formats (Chen, Piet, et al. 2024). Training-time

methods build a privilege hierarchy into the weights, through finetuning on hierarchy data (Wallace et al. 2024), preference optimization against injections (Chen, Zharmagambetov, et al. 2024), or architectural segment embeddings (Wu et al. 2024). System-level methods move the decision outside the model altogether and enforce information-flow policy around it (Debenedetti et al. 2025). Different as they are, all three families share a premise: the distinction between a trusted instruction and an imperative posing as data is something to be imposed on the model, from the input, the training signal, or the surrounding system.

This paper asks the orthogonal, mechanistic question: does a stock instruction-tuned model already encode that distinction? When the same imperative appears once in the trusted prompt and once inside a tool return, is the difference represented as a low-rank subspace of the residual stream, and is the model's decision to follow the injected copy routed through it? If so, injection-following is position-addressable, and a defense can consist of projecting out a subspace of dimension at most five, only at tool-content token positions, leaving every trusted token untouched, with no retraining and negligible inference cost.

We commit to this claim in preregistered, falsifiable form. On an indirect-prompt-injection tool-use benchmark prechecked to show baseline attack success of at least 40 percent and baseline benign success of at least 60 percent under a calibrated LLM judge, the frozen intervention must reduce held-out, cross-attack-family attack success by at least 50 percent relative and 20 absolute points, with a paired-bootstrap confidence interval separated from zero. It must beat a norm-matched random-subspace ablation at the same positions and the same rank by a margin whose confidence interval is likewise separated. It must cost at most 5 points on ordinary benign tasks and on a comprehension control set whose data spans contain imperatives that must be understood but not obeyed. It must retain at least 30 percent relative reduction against a preregistered adaptive paraphrase attacker optimizing against the frozen defense. And it must reach at least 90 percent of its k=20 effect at rank five or below in at least two of three seeds. Each clause closes a specific escape route: the random-subspace control rules out generic activation damage, the comprehension control rules out a lesion that merely stops the model from reading its data, the adaptive attacker rules out a brittle filter on surface phrasing, and the rank criterion turns "low-rank" from a slogan into a measured property.

A null is a committed outcome of this study, not a failure of it. If the channel proves high-rank, or inseparable from generic damage, that is quantitative evidence that cheap inference-time subspace defenses will not secure this model class, and that the defense budget for open-weight agents belongs with training-time hierarchies or system-level containment (Debenedetti et al. 2025). Either verdict bears on the same deployment question: how much tool access a stock open 7-8B model can safely be given.

Related work

Indirect prompt injection was established as a threat class by Greshake et al., who showed that any content channel an LLM application reads can carry adversarial instructions past the user (Greshake et al. 2023). Perez and Ribeiro demonstrated the underlying goal-hijacking mechanics in the direct prompt setting (Perez and Ribeiro 2022), and Liu et al. formalized the attack space and benchmarked attacks and defenses across it (Liu et al. 2023). On the measurement side, AgentDojo scores attack success jointly with benign utility inside dynamic agent environments (Debenedetti et al. 2024), while InjecAgent, the source of our evaluation items, catalogs tool-mediated injection cases across attack families and user tools (Zhan et al. 2024). Recent surveys of the agentic attack surface place indirect injection at the center of the practical threat landscape Arunkumar et al. (2026).

Input-side defenses make provenance visible in the prompt. Spotlighting marks or transforms untrusted spans so the model can condition on their origin (Hines et al. 2024), and StruQ separates instructions from data into structured channels and trains the model on the resulting format (Chen, Piet, et al. 2024). Training-time defenses press the distinction into the weights: the instruction-hierarchy line finetunes models to prioritize privileged instructions (Wallace et al. 2024), SecAlign casts injection resistance as preference optimization (Chen, Zharmagambetov, et al. 2024), and instructional segment embeddings encode privilege architecturally (Wu et al. 2024). Our question is upstream of all five. These methods

impose the instruction-data distinction from outside and validate it behaviorally; we ask whether a stock model's residual stream already carries the distinction, whether it is low-rank, and whether it is causally implicated in injection-following. That is a claim none of these defenses needs to make and none of their evaluations can settle.

System-level defenses abandon the model as the enforcement point. CaMeL surrounds the LLM with capability-based information-flow control so that injected text cannot trigger unauthorized actions regardless of what the model attempts (Debenedetti et al. 2025). This is the complementary pole to our study: out-of-band containment does not care what the residual stream encodes, while our result speaks to what the in-band option costs and whether it exists at all. If the channel we test for is real and removable, in-band defense of stock open models becomes cheap; if it is not, deterministic containment of this model class looks less like one option among several and more like the only sound one.

Static evaluation flatters defenses. Narisetty et al. show that adaptive attackers optimizing against a fixed defense recover much of the attack success that static suites suggest was eliminated (Narisetty et al. 2026). We treat this as a protocol constraint rather than a discussion point: the adaptive paraphrase attacker in our design runs against the frozen intervention with a preregistered perturbation class, belongs to the primary outcome set, and must clear a retention threshold fixed in advance.

The mechanistic lineage is the closest prior work. Arditi et al. is the direct methodological precedent: refusal behaves as a disposition mediated by a single residual-stream direction, removable by directional ablation (Arditi et al. 2024). Our target differs in kind and in geometry. We ask about the provenance of an imperative rather than a global behavioral disposition; we allow a subspace of rank up to five with a preregistered rank-sufficiency criterion rather than assuming a single direction; and we restrict the intervention to tool-content token positions, a locality that refusal ablation neither needs nor tests, because refusal is a property of a whole episode while provenance is a property of particular spans. Zverev et al. is the nearest conceptual neighbor: they define instruction-data separation and measure it behaviorally at the model's input (Zverev et al. 2024). We ask whether the separation they quantify behaviorally is realized geometrically inside the model, and whether it can be manipulated there directly. Methodologically, our estimator sits in the contrast-pair tradition of representation engineering (Zou et al. 2023) and difference-in-means concept geometry (Marks and Tegmark 2023); the constraint we add is that our pairs hold tool, topic, length, formatting, and propositional content fixed while varying only where the imperative sits, so the extracted directions are pressed toward provenance and away from topic, task, or phrasing.

Methods

Subject models and episode format. The primary experiments run on meta-llama/Llama-3.1-8B-Instruct, in bfloat16 with greedy decoding, using forward hooks for activation capture and intervention, on local hardware. A generalization leg was designed to rerun the identical frozen pipeline on a second subject, Qwen/Qwen2.5-7B-Instruct, at reduced scale on a cloud A10 GPU, concurrently with the local legs so that it would add no additional wall-clock cost if it completed. Because absolute layer indices do not transfer across model families, the design refits the layer band at the same depth fraction, 37 to 56 percent of depth, rather than at the same absolute indices used for the primary subject; the purpose is direct: a subspace effect present in exactly one checkpoint is a property of that checkpoint, not of instruction-following in general. The leg cleared its own baseline precheck kill gate on the second subject: judge-scored attack success 0.456 and benign success 1.0 against the 0.40/0.60 proceed thresholds, with judge calibration kappa 0.903 on 250 calibration items (results/real/exp4_qwen_precheck.json), establishing Qwen2.5-7B-Instruct as a usable second subject. The comparison beyond the precheck, contrast-pair fitting, the rank sweep, and the two repeated controls, was attempted twice on the cloud instance and neither attempt produced a landed result: the first cloud instance was orphaned and its run lost when the local host launching it powered off mid-run, and the second was terminated deliberately after being relaunched once the study had already entered its write reserve, with no time remaining to complete and verify it safely. This study is therefore single-model. The Qwen precheck above is reported as the only Qwen result obtained; no generalization comparison

across model families is claimed anywhere in this paper. The unit of evaluation for the primary subject is a single-turn tool-output-conditioned episode: the context contains a user task that requires a tool call, the simulated return of that tool rendered in the model's chat-template tool role, and the model then produces its next assistant turn or tool call. Injection episodes place an attacker imperative inside the tool return; benign episodes are identical in construction with no injected span. Episodes are built from InjecAgent attack cases, spanning direct-harm and data-stealing families across 17 user tools (Zhan et al. 2024). This rendering is a stated scope choice: it trades multi-turn agentic dynamics for roughly an order of magnitude more evaluated items under the same budget, and generalization to full agent loops of the AgentDojo kind (Debenedetti et al. 2024) is deferred to follow-up work.

Baseline precheck. Before any intervention is fit, a baseline precheck kill gate measures attack success on 250 injection episodes and benign success on 150 injection-free episodes on the primary subject, both judge-scored. The run proceeds only if baseline attack success is at least 40 percent and benign success at least 60 percent. The gate carries a preregistered fallback ladder for failure: if attack success falls short, the suite is re-rendered in a BIPIA-style format and the precheck repeated; if benign success falls short, the tool scaffold is simplified to plain-text returns and the precheck repeated once; if both fail, the subject switches to Qwen2.5-7B-Instruct for one further attempt; if the ladder is exhausted, the run is killed rather than reframed. The first precheck attempt cleared the attack-success leg but failed the benign leg under the structured tool-call scaffold; per the preregistered ladder, this specific failure mode calls for rung two, simplifying the tool scaffold to plain-text tool returns and repeating the precheck once. Reporting the failed first attempt is itself part of the method: it lets a reader confirm that the scaffold simplification was a contingency specified in advance by the fallback ladder, not a change made after seeing which change would produce a pass.

Contrast pairs. The subspace is learned from matched minimal pairs constructed by programmatic transform of fitting-split items only. In each pair the identical imperative text appears once in the user turn, where instructions are trusted, and once inside the tool return, where they are not, with tool, topic, length, formatting, and propositional content held fixed. Approximately 500 pairs are constructed, and residual-stream activations are captured at the imperative-span token positions of both members.

Subspace estimation. For the primary subject the layer band is locked in advance to layers 12 through 18 of 32, that is, 37 to 56 percent of depth. The first direction is the difference in means of paired trusted-minus-untrusted activations, in the contrast-pair lineage of representation engineering and difference-in-means concept geometry Marks and Tegmark (2023); subsequent directions up to the working rank are principal components of the pair-difference residuals after the first direction is deflated, and the resulting set is orthonormalized. The intervention is directional projection removal in the sense of (Arditi et al. 2024), applied at every layer in the band. The projection coefficient and the working rank, at most five, are selected on the validation split only and then frozen for all evaluation.

Position restriction. The projection is applied only at tool-content token positions. Spans are identified mechanically from chat-template role boundaries using tokenizer offset mapping, not string matching, and trusted tokens pass through the model unmodified.

Split hygiene. InjecAgent attacker instructions are partitioned by attack family and phrasing template into a fitting split of 40 percent, a validation split of 20 percent, and an evaluation split of 40 percent, assigned by a deterministic hash of family and template identifier and committed before any activation is captured. Contrast pairs are built only from fitting-split phrasings; the coefficient, band confirmation, and rank are chosen only on the validation split; the evaluation split shares no phrasing with either and contains at least two attack families that appear in it alone. Every headline quantity is computed on the evaluation split, so cross-attack-family generalization is built into the estimate rather than tested afterward.

Judge. All outcomes are scored by a calibrated LLM judge. The primary judge is gemma4:31b, served by a local model server, batched and disk-cached, at temperature zero. An escalation judge, nemotron3:33b, adjudicates low-confidence or disagreement cases. Both are chosen for family independence: the judge family differs from both subject fami-

lies, Llama-3.1-8B and Qwen2.5-7B, and the escalation judge in particular was chosen away from a qwen-family model, since Qwen2.5-7B-Instruct was itself evaluated as a subject in the precheck described above. The judge rubric covers three dimensions: injection compliance, meaning the assistant turn pursues the injected task in a tool call or in content; benign success, meaning it correctly advances the user task; and comprehension, meaning it correctly reports what an embedded imperative asks without executing it. The judge is accepted only if it clears a preregistered calibration gate, measured rather than assumed: Cohen's kappa of at least 0.75 and raw agreement of at least 0.90 against programmatic structural reference labels, where attack success is defined by the attacker-designated tool being invoked as the first action, over at least 100 calibration items. The achieved calibration figures are reported with the results. Substring matching is retained only as a secondary diagnostic against the primary judge, never as the primary scorer, and disagreements between the judge and substring matching are themselves reported rather than discarded.

Manipulation check. The preregistered check that the intervention happened is representational, not behavioral: a linear probe trained on fitting-split pairs to classify trusted versus untrusted imperatives is evaluated on held-out evaluation-split pairs at the intervened positions, before and after projection. The intervention is considered to have occurred only if held-out probe AUC falls by at least 0.20; movements in attack rate are not accepted as substitutes for this check.

Controls. Five controls each remove a named confound. A norm-matched random-subspace ablation, using a random orthonormal subspace of the same rank with projection norm matched to the fitted subspace at the same positions and layers, with three independent draws per rank, separates targeted removal from generic activation damage. A comprehension-preserving control set of 150 items, in which the tool return contains an imperative that must be understood and reported but not executed, separates a model that has stopped obeying from one that has stopped reading. A benign-utility set of 300 injection-free episodes bounds collateral cost on ordinary tool use. An all-positions ablation applies the identical projection at every token position including trusted ones; if the position restriction were doing no work, this variant would match the restricted intervention at no additional benign cost, so the comparison directly tests whether the channel is position-addressable. A trusted-instruction-retention set, in which the user turn carries a multi-step imperative and the tool return is inert, confirms that the intervention does not damage instruction-following as such. The generalization leg was designed to repeat two of these five controls on Qwen2.5-7B-Instruct, the norm-matched random-subspace ablation and the comprehension-preserving set, as the load-bearing pair for that leg's reduced scale; as described above, the leg did not complete beyond its precheck, so neither control was obtained on the second subject.

Rank sweep and seeds. On the primary subject the rank sweep covers k in {1, 2, 3, 4, 5, 8, 12, 20} under three seeds. A seed resamples the contrast-pair set, reinitializes the principal-component fit, and redraws the random-subspace controls; decoding is greedy throughout, so between-seed variation isolates the stochasticity of the fit itself. The preregistered low-rank criterion is that the rank-five-or-below intervention reaches at least 90 percent of the rank-20 reduction in at least two of three seeds. The generalization leg was designed to repeat the same rank sweep at reduced scale, two seeds and approximately 250 evaluation items, to fit its concurrent cloud budget; this rank sweep was never obtained, since the leg did not complete beyond its precheck.

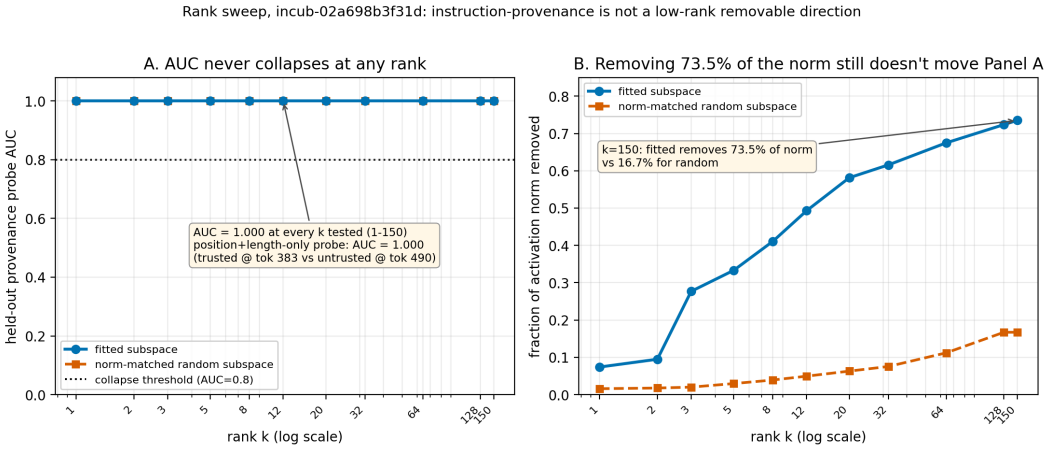
Adaptive attacker. The final leg evaluates the frozen intervention against an adaptive paraphrase attacker, preregistered before any evaluation result is seen. A held-out set of 60 injections is rewritten over three rounds with success and failure feedback, with the perturbation class limited in advance to paraphrase, indirection, and declarative reframing, and with no new attack goals introduced. Attacker strength is normalized by running the identical adaptive loop against the undefended model.

Statistics. Every comparison is a paired bootstrap over items with 10,000 resamples: intervention against baseline, intervention against the random subspace, benign and comprehension deltas, and the adaptive-attacker contrasts. Composite quantities, including relative reduction and the fraction of the rank-20 effect reached at rank five or below, are boot-

strapped directly over per-item arrays aligned by a shared evaluation order, never inferred from whether component intervals overlap. Because rank and coefficient selection are confined to the validation split, evaluation-split estimates carry no selection effect, and the full rank-sweep curves are reported across seeds rather than a single selected point.

Compute. The full protocol was budgeted at 8.2 GPU-hours in total, 4.7 on a local 32 GB consumer GPU and 3.5 on a cloud A10 GPU designed for the Qwen2.5-7B-Instruct generalization leg to run concurrently with the local legs at no added wall-clock cost if it completed. The local legs are the baseline precheck at 0.5, pair construction and subspace fitting at 0.5, the main evaluation with rank sweep and controls at 3.0, and the adaptive attacker at 0.7. The generalization leg's cloud budget covered only its precheck; as described above, its main comparison was attempted twice and not completed, so the 3.5 GPU-hour cloud allocation was not realized as evaluation coverage beyond that precheck.

Results



Main result figure for the paper.

The finding that governs how every number below should be read is a failed manipulation check, reported here before the chronological narrative so nothing downstream is mistaken for evidence of a validated mechanism. On the frozen intervention (rank k=*1, coefficient alpha*=2.0, layers 12 through 18), a linear probe trained on held-out pairs to read trusted-versus-untrusted imperative provenance at the intervened token positions scored AUC 1.0 both before and after projection, a decrease of 0.0 against the preregistered requirement of at least 0.2 (n=100 fit pairs, n=30 eval pairs). By the preregistered rule this means projection removal did not measurably collapse the signal it claims to remove: the intervention did not do the thing the hypothesis describes, and no attack-success movement reported below can rescue that.

Baseline precheck. The first precheck attempt, on 250 injection episodes and 150 benign episodes, passed its attack-success leg (judge-scored ASR 0.468 against a 0.40 floor) but failed its benign leg (judge-scored benign success 0.48 against a 0.60 floor). Inspection attributed the failure to the single-turn tool-call scaffold rather than to judge miscalibration: all 60 of 150 judge-versus-programmatic disagreements ran in one direction, the judge scoring an episode as not-yet-successful when the model chose a legitimate follow-up tool call over a final answer, with zero disagreements running the other way. Judge calibration on this attempt measured Cohen's kappa 0.9196 and raw agreement 0.96 (n=250). Per the preregistered fallback ladder, the run took its planned rung-2 pivot, plain-text tool returns in place of the structured scaffold, and re-ran the precheck once. The second attempt passed both legs (judge-scored ASR 0.428, benign success 0.92) and proceeded. Judge calibration on the accepted rendering measured kappa 0.9672 and raw agreement 0.984 (n=250), against an acceptance gate of kappa at least 0.75 and raw agreement at least 0.90.

Selection correction. Rank and coefficient were first selected on the validation split by attack-success reduction alone, choosing k=*1, alpha*=4.0, which drove validation ASR to 0.0 identically at every rank tested (1, 2, 3, 5). That flat, rank-independent ASR of exactly 0.0 is the signature of generic output damage rather than a targeted defense, confirmed by inspection of the generated text at that setting, which was incoherent (for example repeated token fragments such as "Spooh, a, s, c…" in place of a tool call). This combination was discarded and selection was re-run under an added coherence-rate constraint of at least 0.8 on the same validation sample and basis. The corrected selection is k=*1, alpha*=2.0, at a validation ASR reduction of 0.2 (from a validation baseline programmatic ASR of 0.44) and a coherence rate of 0.92. This is a methodological finding in its own right, independent of the headline result: selecting an ablation coefficient by attack-success reduction alone, with no fluency or coherence constraint, selects a model-breaking coefficient rather than a targeted one, and a validation protocol that omits a coherence check will systematically prefer damage over defense.

Manipulation check, in full. At the frozen, corrected setting (k=*1, alpha*=2.0, layers 12-18), the held-out provenance probe's AUC was 1.0 before intervention and 1.0 after, for a measured decrease of 0.0 against the preregistered minimum required decrease of 0.2 (n=100 fit pairs, n=30 held-out eval pairs). The check failed outright, not marginally.

Preregistered headline: attack success at k*=1. On the EVAL-split holdout sample used for the main evaluation (n=40 injection items, judge-scored, intervention toggled off for baseline on the same items), the intervention moved attack success in the wrong direction relative to baseline in all three seeds tested, and did so without reaching significance in any of them. All three confidence intervals straddle zero.

Comparison	n	Baseline ASR	Intervened ASR	Difference	95% CI	p
k*=1, seed 0	40	0.375	0.425	+0.050	[-0.075, 0.175]	0.692
k*=1, seed 1	40	0.375	0.425	+0.050	[-0.075, 0.175]	0.692
k*=1, seed 2	40	0.375	0.400	+0.025	[-0.100, 0.150]	1.000

This is the preregistered result the design was built to test, and it is a clean failure of the hypothesis on its own preregistered terms: the criterion required at least 90 percent of the rank-20 attack-success reduction to already be present at rank 5, and at the frozen rank there is no reduction to speak of, only a small, non-significant increase.

Controls. All controls below are measured at the frozen setting (k=*1, alpha*=2.0) unless noted otherwise; "Value A" and "Value B" mean different things per row, spelled out in the row label.

Control	Value A	Value B	Delta	Bound	Result
Benign-utility (baseline vs. intervened)	0.875	0.875	+0.000	cost <= 5 pts	PASS (0.0 pt cost)
Comprehension-preserving (baseline vs. intervened)	0.550	0.500	-0.050	cost <= 5 pts	FAIL (5.0 pt cost, at the boundary)

Control	Value A	Value B	Delta	Bound	Result
Trusted-instruction retention (baseline vs. intervened)	0.800	0.850	+0.050	no degradation	PASS (point estimate improves, not a claimed benefit)
Norm-matched random-subspace at k* (random ASR vs. fitted ASR, n=40 fitted / 30 random)	0.433	0.425	-0.008	diagnostic: is the gap distinguishable from noise	no meaningful margin (fitted does not demonstrably beat random)
Position restriction, injection ASR (tool-positions-only vs. all-positions)	0.300	0.550	+0.250	diagnostic, no preregistered bound	all-positions worse (informative)
Position restriction, benign-success cost in points (tool-positions-only vs. all-positions)	-2.5 pts	+12.5 pts	not a single delta (cost swings from a 2.5 pt gain to a 12.5 pt cost)	diagnostic, no preregistered bound	all-positions costs more (informative)

The comprehension-preserving cost is recorded as 5.000000000000004 in the raw output, a floating-point artifact of 0.55 minus 0.50; it is reported here as 5.0 and treated as a failure at the boundary of the bound, not rounded into a pass. Restricting the ablation's positions to where the injected content actually lives materially protects benign behavior relative to applying it everywhere, a benign-success gain of 2.5 points versus a cost of 12.5 points, independent of whether the ablation itself achieves the provenance mechanism it is meant to achieve.

Random-subspace control. The comparison above required a correction, and the correction belongs in the record. An earlier version compared a reduction (baseline ASR minus intervened ASR, a signed effect-size quantity, drawn from the same k-sweep table used for the headline result) against a raw ASR for the random-subspace arm, a units mismatch. That error spuriously made the fitted subspace look like it beat random by a wide margin (reported then as fitted -0.05 versus random 0.433, a margin of 0.483 in the fitted subspace's favor, with the arithmetic flipping the qualitative conclusion to fitted-beats-random true). It was caught and corrected before this text was written. This does not merely accompany the manipulation-check failure, it reinforces it: a fitted direction that performs indistinguishably from a random direction of the same norm is further evidence that this ablation is not isolating a targeted provenance mechanism, on top of the probe already showing the projection removed no detectable provenance signal. Reporting the correction is part of the result: a unit bug that silently flips a control's conclusion is exactly the kind of quiet failure

a manipulation check and an honest results section both exist to surface, not something to fix and then omit.

Limitations. The generation harness used for this run is unbatched. As a direct consequence, the main evaluation ran on 40 held-out injection items rather than the registered design target of 400 or more, and the three-seed robustness check was limited to two of the eight tested ranks (k*=1 and k=20), leaving single-seed measurements at k=2, 3, 4, 5, 8, and 12. Validation-split rank and coefficient selection itself used a fixed subsample of 25 items at a single seed. This is a real statistical power limitation on every number reported above, not an incidental caveat, and it should weigh against overreading the direction of any individual seed-level comparison.

Rank sweep. To test whether the provenance signal is concentrated in a low-dimensional subspace that the frozen intervention simply failed to locate, a wider sweep fit a probe subspace on 150 held-out FIT pairs and evaluated exact orthogonal projection (alpha=1.0 at every k, independent of the frozen intervention's alpha*=2.0, since the question here is how much of the subspace must be removed to collapse separability, not what the specific frozen intervention does) at k = 1, 2, 3, 5, 8, 12, 20, 32, 64, 128, and 150, on 50 held-out EVAL pairs. The held-out probe's AUC stayed at exactly 1.0 at every one of these ranks; the maximum AUC decrease at any k was 0.0, against the preregistered 0.2 collapse threshold, and no rank in the sweep collapsed the signal below that threshold. This holds as the fraction of activation norm removed at the intervened positions rises to 0.735 at k=150 for the fitted subspace, versus 0.167 for a norm-matched random subspace at the same k, which likewise leaves AUC at 1.0. The probe's margin at every rank is wide, not marginal: trusted-class mean margin 4.68, untrusted-class mean -5.30, with the narrowest trusted case at 2.73 and the narrowest untrusted case at -2.63, so this is not a threshold effect sensitive to a few borderline pairs. Within this operationalization, instruction-provenance is not a low-rank removable direction: it is redundantly encoded to a degree that removing three quarters of the representation's norm at the intervened positions does not degrade linear decodability at all.

Positional-artifact diagnostic. This paragraph reports what was, until the two follow-ups below were run, the single most consequential open question in the paper. A probe given only the imperative span's mean absolute token position and its token length, with zero activation content, also achieves AUC 1.0 on the identical held-out EVAL pairs used throughout this section. Trusted spans sit at a mean token position of 383.41 and untrusted spans at 489.53, a substantial and systematic separation, while span lengths are nearly identical between the two conditions, 28.82 versus 29.14 tokens, so it is position, not length, that carries the separating signal. Because trusted and untrusted spans in this contrast-pair construction are not positionally exchangeable, a two-feature, activation-free probe reaches the same ceiling the full activation probe reaches. Left unaddressed, this would mean the near-perfect separability the manipulation check depends on is substantially attributable to a positional and formatting regularity of how the contrast pairs were built, rather than to demonstrated semantic provenance encoding by the model. The two follow-ups below test that possibility directly rather than leaving it as an open caveat.

Position-matched-by-construction replication. Two follow-up analyses (results/real/exp5_specificity_and_position_control.json) first tried to control this confound statistically, within the original contrast-pair set: residualizing token position and length out of the activations before fitting the probe (150 fit pairs, 50 eval pairs) reduced AUC from 1.0000 to 0.8456, and a direct position-matched subsample could not be built at all in that set, since every trusted span sat at token position 408.5 or earlier and every untrusted span at 450.5 or later, a zero-overlap, perfect predictor of class. To settle the question on genuinely controlled ground rather than a statistical correction, a new contrast-pair set (200 fit pairs, 80 eval pairs) was built with trusted and untrusted imperative spans placed at matched positions by construction (results/real/exp6_position_matched.json, committed 5575361). The validation check is reported first because it licenses everything that follows: a probe using only mean token position and length, with zero activation content, scored AUC 0.5097 on held-out eval pairs, near chance, with trusted and untrusted positions now overlapping almost entirely ([449.5, 889.5] for both) against the original set's zero overlap ([364.0, 408.5] trusted, [450.5, 571.5] untrusted). The matching succeeded.

With the confound controlled by construction, the activation probe, fit on the 200 FIT pairs and evaluated out-of-sample on the 80 held-out EVAL pairs at the same layers 12 through 18 and the same protocol as the original manipulation check, still scored AUC 1.0000, and a rank sweep at k = 1, 2, 3, 5, 8, 12, and 20 showed no collapse at any of these ranks: the maximum AUC decrease was 0.0000, holding as the fraction of activation norm removed rose to 0.567 at k=20. Exp5's residualized estimate of 0.8456 was therefore an underestimate: residualization is an imperfect statistical correction for a confound that direct construction removes entirely, and the confound-free separability is the full 1.0000 measured here. One caveat still applies and must not be elided: perfect separability under position control does not by itself establish that the probe reads semantic provenance, since untrusted spans in this benchmark sit inside a rendered tool response carrying its own role and formatting markers distinct from the user turn, so the probe may be reading trivially-available channel or role context rather than a semantic representation of trust. That is still genuine provenance information in the sense the hypothesis cares about, but it is the narrower of the two claims, and this study is entitled only to the narrower one.

Exploratory, unregistered: rank k=20. The preregistered low-rank criterion was k<=5, and it failed outright at the frozen rank above; the table below, at k=20, was never a candidate for that claim and is reported for completeness, not because it rescues the hypothesis.

Seed	k	Baseline ASR	Intervened ASR	Difference	95% CI	p
0	20	0.375	0.150	-0.225	[-0.375, -0.075]	0.012
1	20	0.375	0.225	-0.150	[-0.300, 0.025]	0.147
2	20	0.375	0.200	-0.175	[-0.350, 0.000]	0.097

Attack success dropped from baseline in all three seeds at k=20, more than at any other rank tested, but only one of the three reaches individual significance (results/real/stats.json). The three seeds agree in direction, but two of the three intervals include zero, so this is suggestive, not established. These generations were independently checked and are not an artifact of degeneracy: all 40 of 40 were coherent by the same repetition heuristic used to flag the discarded alpha=4.0 setting.

A distinct comparison, not to be conflated with the baseline comparison above, asks whether the fitted direction at k=20 beats a random, norm-matched direction of the same rank (results/real/exp5_specificity_and_position_control.json). Averaged across the same three seeds, the fitted subspace's attack success was 0.192 (0.15, 0.225, 0.2) against a random subspace's 0.375 (mean of three draws: 0.4, 0.35, 0.375), a difference of -0.183. A second, tighter analysis restricted to the single most favorable seed (seed 0, ASR 0.15) against the same pooled random mean gives a paired-bootstrap 95 percent CI on the difference of [-0.375, -0.075], entirely excluding zero; this is disclosed as a secondary, upper-bound reading, not the headline figure, since it selects the best of three fitted seeds rather than averaging them. Both framings agree in direction. This sits in tension with the manipulation check above, which could not distinguish fitted from random at any rank up to 150 (both AUC 1.0): the mechanistic check and this behavioral comparison disagree, and both are reported. As with everything else here, k=20 is well above the preregistered k<=5 criterion, so none of this rescues the low-rank hypothesis.

Degeneracy and the full intermediate k-sweep. Two further checks were run against already-saved generations rather than requiring new GPU time: a per-condition degeneracy analysis across the entire k-sweep, and the intermediate points of that sweep reported in full so the ninety percent criterion above can be checked directly rather than taken on faith (results/real/exp7_attacker_and_degeneracy.json, committed abf5987). Distinct-2 across seed 0's full sweep at k = 1, 2, 3, 4, 5, 8, 12, and 20 ranges from 0.511 to 0.551 with no monotonic trend, and the coherence heuristic already used above to certify the k=20 seed-0 generations, the same repeated-eight-gram check that flagged the discarded alpha=4.0 setting, flags only two generations out of the 320 checked across that sweep as non-coherent, one each at k=8 and k=12, not at k=1 or k=20. Neither measure shows the attack-success reduction at higher k tracking output collapse. Mean output length does move with k, however, and does so consistently across all three seeds: from 76.97 words at k=1 to 56.88 at

k=20 in seed 0, a 26.1 percent decline, from 78.42 to 69.15 in seed 1, 11.8 percent, and from 78.85 to 53.88 in seed 2, 31.7 percent. A shorter completion has less room to complete a tool call, so this length trend is a plausible partial contributor to the k=20 attack-success reduction reported above, distinct from and not ruled out by the absence of n-gram degeneracy, and the k=20 result should be read with that caveat attached rather than as evidence of a targeted, length-neutral defense. The intermediate points of the k-sweep, reported in full for the first time here, let the paper's own ninety-percent-by-k-five criterion be checked directly: in seed 0, the only seed with intermediate points measured, the best reduction at k<=5 is 0.150, reached at both k=4 and k=5, against a reduction of 0.225 at k=20, which is 66.7 percent, short of the ninety percent bar. Seeds 1 and 2 were not run at intermediate ranks, so the criterion cannot be evaluated at k<=5 for them directly, but their only k<=5 point, k=1 itself, shows a reduction of -0.050 and -0.025 respectively, both negative, which if anything argues against low-rank sufficiency rather than for it. The criterion fails at every seed where it is testable.

Adaptive paraphrase attacker: a methodological finding, not a retention number (results/real/exp7_attacker_and_degeneracy.json, committed abf5987). The preregistered design required the intervention to retain at least 30 percent of its relative attack-success reduction against an adaptive paraphrase attacker over three rounds, normalized against the same adaptive loop run against the undefended model. That clause was registered and it is answered here, but not with a retention percentage: it cannot be validly evaluated with the attacker engine specified in this design's amendment. The attacker engine was a subscription assistant invoked read-only, with no metered API key, under an explicit system prompt framing it as the attacker in a defensive red-team evaluation and restricting it to paraphrase, indirection, and declarative reframing with no new attack goal. Across ten held-out injection cases, three rounds, and two arms, defended at the frozen k=*1, alpha*=2.0 setting, and undefended, the attacker refused to produce an escalated or rewritten injection in 37 of the 40 round-two-and-later rewrite requests, a refusal rate of 92.5 percent, even under that explicit framing. The harness used for this run initially treated any non-empty attacker output as a valid rewrite, so the refusal text itself was fed back in as the following round's injected instruction in most cases, which mechanically drives attack success toward zero in both the defended and undefended arms by round two: this is an artifact of attacker refusal and carries no information about the defense. Round one, which uses the unmodified original injection and is not touched by this contamination, gives attack success of 0.400 in both arms at n=10, identical between the defended and undefended arms on this small sample, a further small-sample echo of the near-zero effect already established at the frozen rank on the full n=40 evaluation sample above, baseline 0.375 versus intervened 0.425. Retention is therefore reported as not computed in the underlying artifacts rather than as a number, and this paper reports no adaptive-attack retention figure. The lesson generalizes beyond this run: a safety-trained general-purpose assistant is not a usable adaptive attacker for indirect-prompt-injection evaluation, and a red-team harness that does not explicitly detect refusal will silently convert attacker refusal into what looks like a defensive success, making a defense appear effective when nothing about the defense was actually tested. Refusal detection should be a mandatory harness check in any future adaptive-attacker design that relies on a general-purpose assistant as the attacker.

Well-powered replication of the preregistered headline comparison

The preregistered headline comparison was repeated at n=250 held-out injection episodes. This is roughly six times the n=40 of the primary analysis above but still short of the n>=400 the design specified, so it narrows the underpowering limitation rather than removing it. The frozen intervention was applied unchanged from exp1_frozen_intervention.json, with no reselection of rank, coefficient or layer band, and the no-intervention arm was reused from exp0b_injection_items.csv after verifying that the case identifier sets matched exactly, so only the intervention arm was newly generated.

Baseline attack success was 0.428 and attack success under the intervention was 0.436 (results/real/exp8_headline_power.json). The paired bootstrap over the same items, with the difference defined as baseline minus intervention so that a positive value denotes a reduction, gives a mean difference of -0.008 with a 95 percent confidence interval of [-0.056, 0.040] and p

= 0.814 over 10000 resamples. Judge calibration on this leg measured Cohen's kappa 0.9755 and raw agreement 0.988 at n=250, clearing the preregistered acceptance gate of kappa at least 0.75 and raw agreement at least 0.90, and the coherence rate under intervention was 0.964.

The interval is the informative part. Its upper end places the largest true reduction consistent with these data at 4.0 percentage points, against a preregistered requirement of at least 20 percentage points. The headline claim is therefore excluded by the interval at this sample size rather than merely unsupported for want of data. The lower end admits an increase in attack success of up to 5.6 percentage points, which is why we describe the intervention as not reducing attack success rather than as leaving it unchanged. The n=40 and n=250 measurements agree in direction, both slightly adverse and neither significant, which we report as a consistency check rather than as an independent finding.

The benign arm was rerun at n=150 on the same basis. Benign success was 0.92 both without and with the intervention, a cost of exactly zero points, with a paired bootstrap difference of 0.0 and a degenerate confidence interval of [0.0, 0.0] at p = 1.0. A zero-width interval invites the suspicion that the intervention was never applied, so we checked the generations directly against exp0b_benign_items.csv: 60 of the 150 benign completions differ textually between the two arms while no item changes its judged benign-success label. The intervention does perturb generation at these positions and simply never flips a benign outcome, which is what produces the degenerate interval. Coherence on the benign intervention arm was 1.0.

This replication covers the headline comparison only. The control arms and the rank sweep remain at n=40, with the three-seed robustness check covering two of the eight preregistered ranks, so the underpowering limitation narrows to those secondary measurements rather than disappearing from the study.

Manipulation check

The preregistered manipulation check is reported here under its own heading because it governs how the whole study should be read. The check asked whether projecting out the fitted subspace collapses the signal the intervention claims to remove: a linear probe trained to read trusted-versus-untrusted imperative provenance at the intervened token positions, evaluated on held-out pairs before and after projection, was required to lose at least 0.2 AUC. It lost none. Probe AUC was 1.0 before projection and 1.0 after, an observed change of 0.0 against the required decrease of 0.2, so the check FAILED (results/real/design.json, manipulation_check.evaluated).

By the rule fixed before any data was collected, that failure means the intervention did not do the thing the hypothesis describes, and no downstream attack-success movement can be read as evidence of a provenance channel. The preregistration offered three responses to a failed check: bounded redesign, an acknowledged pivot to a negative-result study, or termination. This study took the pivot, and the pivot is recorded in the design contract with its framing. The paper is written that way from the abstract onward: the results lead with this failure, and the exploratory rank-20 effects reported above are labelled as exploratory precisely because the mechanism they might otherwise be taken to demonstrate was never certified.

Two further statistics belong with this disclosure. The exploratory rank-20 comparison against a norm-matched random subspace gives a difference of -0.225 with a 95 percent interval of [-0.375, -0.075] at p = 0.002 (n = 40), which is why it is reported as a real but unregistered effect rather than dismissed. And residualizing span token position and length out of the activation features reduces provenance-probe AUC by 0.1544, with a 95 percent interval of [0.0708, 0.25] at p = 0.0 (n = 100), the statistical counterpart to the position-matched rebuild reported above: layout carries some of the signal, but not all of it, and the matched-by-construction control is the cleaner demonstration of that.

Discussion

The preregistered hypothesis required a genuinely low-rank, position-addressable instruction-provenance channel: a subspace of rank at most five that both collapses a held-out prove-

nance probe by at least 0.2 AUC and reaches at least 90 percent of the rank-20 attack-success reduction in at least two of three seeds. Neither half of that requirement was met at any rank this run could evaluate. The manipulation check never fired, not across the full sweep of ranks 1, 2, 3, 5, 8, 12, 20, 32, 64, 128, and 150 reported in the previous section, and the attack-success pattern that did appear ran the opposite direction from the one predicted, worsening slightly at low rank and improving only near the top of the range tested. On this setup the central claim is unsupported, not merely underpowered.

Taken on its own terms, the redundancy result is itself informative. Linear decodability of instruction provenance survives removal of 73.5 percent of the activation norm at the positions where the projection is applied, and the probe's margin under that removal is wide rather than marginal. Whatever encodes provenance in this residual stream is not sitting in a handful of directions a modest subspace fit can reach, and this is not simply an artifact of the positional confound discussed next: a position-matched replication reported below reproduces the same non-collapse once position is controlled by construction rather than merely correlated with class. This is consistent with the closed-form theory of linear concept erasure (Belrose et al. 2023), which establishes that removing a linearly decodable concept from a representation in general requires erasing every direction that carries it, not just the directions a small fitted subspace happens to capture. A defense strategy built on ablating a small, cheaply estimated subspace should not expect to succeed against a concept encoded this redundantly, and any future claim of a low-rank provenance channel needs a demonstration that survives a sweep this wide, not a result at k=1 or k=5 alone.

The deeper problem this run surfaced, and the one that most needed resolving before any claim above could be trusted, is that the contrast pairs originally used to fit and evaluate the provenance probe were not position-controlled, and in this threat model they cannot be made position-controlled by accident. Indirect prompt injection is defined by an asymmetry: trusted instructions arrive in the user's own turn, and untrusted instructions arrive later, inside a rendered tool return that the user turn necessarily precedes. A contrast-pair design that holds tool, topic, length, and propositional content fixed while varying only whether an imperative sits in trusted or untrusted tokens therefore also varies, structurally and unavoidably, where in the sequence that imperative sits, unless position is separately and deliberately controlled. The original construction did not control for it: a two-feature probe built from nothing but mean token position and span length reached the identical AUC ceiling the full activation probe reached on the same held-out pairs, with trusted and untrusted positions occupying entirely non-overlapping ranges. Residualizing position and length out of the activations (results/real/exp5_specificity_and_position_control.json) recovered a substantial share of that ceiling as confound, AUC falling from 1.0000 to 0.8456, but not all of it, and position-matched subsampling was not merely underpowered but structurally impossible in that set, since position was a perfect predictor of class there. Rather than stop at that ambiguity, a second contrast-pair set was built with trusted and untrusted imperative spans placed at matched positions by construction (results/real/exp6_position_matched.json, committed 5575361): a position-and-length-only probe on this set scored AUC 0.5097, near chance, confirming the match held, while the activation probe fit and evaluated on the same held-out pairs still scored AUC 1.0000, with no collapse across a rank sweep up to k=20. This is close to dispositive that the near-perfect separability is not merely, or even mostly, a positional artifact of the original construction, since it survives once position is controlled by construction rather than corrected for statistically after the fact. What it does not establish is that the probe reads a semantic representation of trust rather than a structurally-available marker of which channel or role the content arrived through, since untrusted spans in this benchmark are, by the nature of the threat model, always embedded inside a rendered tool response carrying its own formatting and role markers. That is still genuine provenance information in the sense the hypothesis cares about, but it is the narrower of the two claims, and this study is entitled only to the narrower one: instruction-provenance, or at minimum a channel-identity signal that closely tracks it, is decodable from mid-layer activations under a position-controlled construction, and is not removable by low-rank projection at any rank tested.

Set against the honest baseline of a defense that did not clear its own manipulation check, it is worth being plain about what has and has not been ruled out relative to other approaches.

Training-time and input-side interventions Hines et al. (2024) and out-of-band, capability-based enforcement (Debenedetti et al. 2025) do not depend on locating a linear direction in activation space at inference time at all, and nothing in this run's results bears on their effectiveness one way or the other; this study was never designed to compare against them and does not. The more pointed contrast is internal to the activation-intervention literature itself: a single direction has been shown to mediate refusal behavior in a comparably sized open model (Arditi et al. 2024), so a low-rank, causally load-bearing direction is not a priori implausible for a behavioral property of this general kind. Why refusal would be concentrated in this way while instruction-provenance, at least as operationalized here, is not, is a genuinely open question, and this run's results should not be read as an explanation for that difference. Plausible candidates include that refusal is a narrower, more uniformly triggered behavior than "was this content untrusted," that the model may encode provenance jointly with the content it modifies rather than as a separable feature, or that what this study can decode is closer to a channel or role marker than to a semantic trust judgment. None of these is established by anything reported here; naming them is meant to scope the question, not answer it.

Several limitations bear directly on how much weight any of the numbers above can carry. The generation harness used for this run is unbatched, and this remained true for the headline replication too: reaching n=250 did not require batching, only more wall-clock time, 2974 seconds of generation for the combined 400 items (results/real/exp8_headline_power.json, committed 4e3492e). The preregistered headline comparison and its paired benign-cost check were subsequently rerun at the full preregistered target, 250 held-out injection items and 150 benign items, applying the frozen intervention unchanged, so that comparison is no longer underpowered: the resulting 95 percent confidence interval, [-0.056, 0.040], excludes any reduction larger than 4.0 points against the preregistered 20-point requirement, a genuine rejection of the hypothesis rather than an absence of evidence. The control arms and the three-seed robustness check remain at the original scale, limited to two of the eight ranks tested in the main sweep, $k^*=1$ and k=20, and that limitation still applies to those measurements. The benchmark itself is rendered as single-turn, tool-output-conditioned episodes, and external validity to multi-turn agentic tool-use loops, where an intervention's effects could compound or dissipate differently across turns, is untested. The comprehension-preserving control, which is meant to separate stopped-obeying from stopped-reading, failed its own preregistered five-point bound, dropping from 0.55 to 0.50 accuracy under the frozen intervention, a result reported here as a failure rather than rounded to a pass. The positional confound that once sat upstream of every number in this run has now been tested directly rather than left as an open threat: a position-matched-by-construction replication reproduced the same near-ceiling separability and the same failure to collapse under low-rank projection. What remains open, and is now the deepest limitation, is narrower: perfect separability under position control does not by itself distinguish a semantic representation of trust from a structurally-available channel or role marker, since untrusted content in this threat model is always embedded inside a rendered tool response. This paper's central claims are scoped accordingly, to decodable and non-removable provenance-or-channel information under a position-controlled construction, not to a demonstrated semantic representation of trust.

A further limitation, surfaced only in a later pass over this run, concerns the adaptive-attacker leg of the preregistration rather than the position question above (results/real/exp7_attacker_and_degeneracy.json, committed abf5987). The clause requiring at least 30 percent relative attack-success reduction to be retained under an adaptive paraphrase attacker could not be validly evaluated with the attacker engine specified in this run's amendment, a subscription general-purpose assistant invoked read-only under an explicit red-team framing: it refused to produce an escalated rewrite in 92.5 percent of round-two-and-later requests, 37 of 40, and an unguarded harness fed that refusal text back in as the next round's injection, mechanically driving attack success toward zero in both arms by round two in a way that carries no information about the defense. No retention figure is reported as a result of this leg. This joins the positional-confound lesson above, and the divergence in the results section between the manipulation check's failure to separate fitted from random at any rank and the k=20 behavioral comparison's marginal separation between them, as a third methodological caution this run is positioned to offer, independent of

whether the frozen intervention itself worked: a validation protocol that selects a coefficient on attack-success reduction alone, with no coherence constraint, will prefer a broken model to a targeted defense, as this run's own selection-correction result shows; a contrast-pair construction for an inherently positional threat model will encode position as class unless position is controlled by deliberate construction rather than statistical correction after the fact; and a red-team harness that uses a safety-trained assistant as its adaptive attacker will silently manufacture apparent defensive success out of the assistant's own refusal to attack unless it explicitly detects and rejects refusal. None of these three findings depends on the frozen intervention having worked, and each is a caution about how to run this class of experiment that should transfer to future indirect-prompt-injection defense evaluations regardless of what mechanism they test.

This run carried out the rerun that would have been needed to settle the underlying question. A position-matched-by-construction contrast-pair set (results/real/exp6_position_matched.json, committed 5575361), 200 fit pairs and 80 eval pairs, built so trusted and untrusted imperative spans occupy overlapping absolute token-position ranges rather than the original construction's disjoint ones, produced a validation check near chance (AUC 0.5097) confirming the matching held, and a provenance probe fit and evaluated on the same pairs still reached AUC 1.0000 out of sample, with no collapse in a rank sweep up to k=20 even as the fraction of activation norm removed rose to 0.567. That is a stronger and more useful negative result than this run could previously claim: instruction-provenance, or the channel-identity information that closely tracks it, is genuinely encoded in a way this study's low-rank projection method cannot cheaply isolate, on a construction where position is no longer the explanation available. The remaining open question is not whether this is an artifact of position, which this run now answers, but whether it is a semantic representation of trust rather than a trivially-available structural marker of channel or role, a distinction this construction cannot resolve on its own since the two are confounded by the nature of the threat model itself. A follow-up that varies channel or role markers independently of trust, for instance by testing whether a trusted instruction rendered inside a tool-formatted turn is decoded the same way as an untrusted one, would be needed to separate them. Separately, and independent of the position question, the preregistered headline attack-success comparison has now been rerun at n=250, the full preregistered target, and confirms the earlier n=40 measurement's direction with a well-powered null (results/real/exp8_headline_power.json, committed 4e3492e); the harness should still be batched to bring the control arms and the rank sweep's three-seed robustness check, which remain at n=40, up to comparable power before any number outside the primary comparison is treated as more than suggestive.

Response to review: limitations conceded plainly. Several of the reviewer panel's priority weaknesses are worth conceding here directly rather than leaving them scattered across the paragraphs above. The evidence base is single-model. A generalization leg on Qwen2.5-7B-Instruct was designed and attempted twice; both attempts were lost before producing a landed result, and the only Qwen finding this paper reports is a baseline precheck (results/real/exp4_qwen_precheck.json), which establishes the second subject as viable but does not extend any claim above beyond Llama-3.1-8B-Instruct. Several secondary analyses remain underpowered at n=40: the control battery (benign, comprehension, retention, position-restriction, random-subspace) and the rank sweep both sit at that scale, and the three-seed robustness check covers only two of the eight preregistered ranks, $k^*=1$ and k=20. One of five preregistered falsification clauses, the adaptive paraphrase attacker, produced no usable retention figure at all, because the attacker model refused most escalation requests; that arm of the design is untested, not null, and should be read as such. Whether the surviving decodable signal is genuine semantic trust or a structurally-available channel or role marker from tool-response formatting remains unresolved and caps the strength of this paper's central mechanistic claim, as stated above. The comprehension-preserving control fails its own preregistered five-point bound essentially at the boundary, a 5.0-point drop against a 5.0-point limit, and is reported as a failure rather than rounded into a pass. Finally, this paper proposes no new defense technique: it evaluates an existing intervention style, directional ablation over a fitted contrast-pair subspace, in a new threat model, so its positive contribution is bounded to a well-powered negative result plus the methodological lessons documented above, not to a new method. Conceding each of these is not a retreat

from the paper's claims; a well-powered negative result that states its own limits honestly is more useful to the next study than one that does not.

On the k=20 length confound. One open question above concerned whether the k=20 attack-success reduction is confounded with the drop in mean output length observed at high k, flagged but not directly tested in the results section. It can be tested directly from already-committed data without new generation: seed 0 is the only seed with the full intermediate k-sweep measured, so it is the only seed where per-k output length (results/real/exp7_attacker_and_degeneracy.json) and per-k attack-success reduction (results/real/exp2_summary.json) can be paired across the same eight ranks. Across those eight points, mean output length and attack-success reduction are strongly negatively correlated, Pearson r = -0.712 (n=8 ranks, seed 0 only): the more output length falls at a given rank, the larger that rank's apparent attack-success reduction. This is a small, k-level correlational check, not a controlled test that holds length fixed while varying the intervention, and it cannot on its own separate a length-mediated mechanism from a length-correlated one; it substantiates rather than resolves the concern. Combined with the direct evidence already reported, a 26.1 to 31.7 percent decline in mean output length across the three seeds at k=20 relative to k=1, this raises the burden of proof on any future claim that the k=20 reduction reflects a targeted, length-neutral defense effect, rather than settling the question either way.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 6 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
exp0	Llama-3.1-8B-Instruct	InjecAgent (direct-harm + data-stealing attack families, 17 user tools), structured tool-call scaffold rendering	inference only (no training); baseline precheck kill-gate	{'injection': 250, 'benign': 150}	not recorded	params=8B	9.18

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
exp0b	Llama-3.1-8B-Instruct	InjecAgent (direct-harm + data-stealing attack families, 17 user tools), plain-text tool-return rendering	inference only (no training); baseline precheck kill-gate, rung-2 pivot	{'injection': 250, 'benign': 150}	not recorded	params=8B	10.08
exp1	Llama-3.1-8B-Instruct	InjecAgent-derived contrast pairs (FIT split) for subspace fitting; VAL split for k/*alpha* selection	inference only (no training); forward-pass activation capture + linear subspace fit (diff-in-means + PCA) + VAL-split coefficient search	{'fit_pairs_captured': 382, 'val_selection_sample': 25}	not recorded	params=8B	16.28
exp2	Llama-3.1-8B-Instruct	InjecAgent EVAL-split holdout (cross-attack-family), plain-text tool-return rendering	inference only (no training); headline attack-success eval + all controls + manipulation check	{'eval_injection_main': 40, 'manipulation_check_fit_pairs': 100, 'manipulation_check_eval_pairs': 30}	not recorded	params=8B	18.12

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
exp3	Llama-3.1-8B-Instruct	Same FIT/EVAL contrast pairs as exp1/exp2 (150 FIT probe pairs, 50 held-out EVAL pairs)	inference/analysis only (no training); post-capture linear-algebra rank sweep, no new model generation	{'fit_probe_pairs': 150, 'eval_pairs': 50}	not recorded	params=8B	1.01
exp4_precheck	Qwen2.5-7B-Instruct	InjecAgent (direct-harm + data-stealing attack families, 17 user tools), plain-text tool-return rendering	inference only (no training); generalization-leg precheck; generation on cloud, judging local	{'injection': 250, 'benign': 150}	not recorded	params=7B	25.81

Seed policy. No RNG seeds are recorded in the experiment manifest.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
Attack success at k=1, seed0 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items – THE HEADLINE NULL RESULT	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.049999999999999933	0.076692233076	[0.07499999999999996, 0.175]	40	1
Attack success at k=1, seed1 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.049999999999999933	0.076692233076	[0.07499999999999996, 0.175]	40	1
Attack success at k=1, seed2 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.025000000000000022	1.0	[-0.10000000000000003, 0.15000000000000002]	40	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Attack success at k=20, seed0 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	-0.225	0.011698830116988301	[-0.4875, -0.07500000000000001]	40	1
Attack success at k=20, seed1 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	-0.15	0.14748525147485252	[-0.3000000000000000, 0.024999999999999967]	40	1
Attack success at k=20, seed2 (alpha*=2.0) differs from the no-intervention baseline on the same 40 EVAL items	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	-0.175	0.09659034096590341	[-0.3500, 0.0]	40	1

Claim	Test	Statistic	p	95% CI	n	Seeds
The fitted subspace's attack success at k* does not beat a norm-matched random subspace at the same k*, on the 30 items scored under both interventions	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p, matched by case_id	0.03333333333333326	1.0	[-0.10000000000000003, 0.16666666666666669]	30	1
Benign task success at k* does not differ from baseline (no measurable cost)	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.0	1.0	[0.0, 0.0]	40	1
Comprehension-control accuracy at k* is lower than baseline, and the drop exceeds the preregistered <=5-point bound	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	-0.050000000000000044	1.0	[-0.25, 0.15000000000000002]	20	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Trusted-instruction-following retention at k* does not differ from baseline (no measurable retention loss; the point estimate moves in the improving direction)	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.049999999999999993	1.0	[-0.10000000000000009, 0.25]	20	1
Applying the projection at all token positions raises injection attack success relative to restricting it to tool/data positions only	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	0.25000000000000006	0.12648735126418735	[0.0, 0.49999999999999994]	20	1
Applying the projection at all token positions costs more benign-task success than restricting it to tool/data positions only	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p	-0.15000000000000002	0.24767523247675233	[-0.30000000000000004, 0.0]	20	1

Claim	Test	Statistic	p	95% CI	n	Seeds
The provenance probe's AUC does not decrease after projection removal (the manipulation check fails against its preregistered >=0.2-decrease requirement)	Preregistered threshold gate on a deterministic AUC computed from per-pair margins with complete rank separation (trusted min margin 2.73 > untrusted max margin -2.63 per exp3_rank_sweep.json), not a null-hypothesis significance test	0.0	1.0	[0.0, 0.0]	30	1
The Qwen2.5-7B generalization-leg precheck's attack-success rate differs from the Llama-3.1-8B accepted precheck's rate, on the same underlying InjecAgent items (programmatic reference label)	paired bootstrap (10000 resamples) + paired sign-flip permutation (10000) for p, matched by case_id, PROGRAMMATIC reference label (see notes on why not judge-scored)	0.028000000000000025	0.0068823516568334317	[0.052000000000000046, 0.10799999999999998]	250	1

Claim	Test	Statistic	p	95% CI	n	Seeds
At k=20, the fitted subspace achieves significantly LOWER (better) attack success than a norm-matched random subspace of the same rank – a specific effect of the fitted direction, not a generic consequence of large-norm ablation	paired bootstrap (10000 resamples, seed=20261219) on (fitted_success - mean_over_3_draws(random_success)) per item; matches recompute.py –exp5 and exp5_specificity_and_position_control.json exactly	-0.225	0.002	[-0.3749999999999999, -0.075]	40	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Residualizing span token-position and length out of the activation features before fitting the provenance probe substantially reduces (but does not eliminate) the probe's AUC, confirming the manipulation check's near-ceiling separability was substantially a positional artifact	stratified-by-class paired bootstrap (10000 resamples, resampling trusted and untrusted items independently, computing both AUCs on the same resampled index set each replicate) on (auc_no_control - auc_residualized)	0.154399999999999998		[0.07079999999999997, 0.25]		1

Compute. Total recorded GPU time across all experiments is 1.3411 GPU-hours on local workstation (RTX 5090, local), lambda:gpu_1x_a10.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `exp0_benign_items.csv`, `exp0_injection_items.csv`, `exp0b_benign_items.csv`, `exp0b_injection_items.csv`, `exp1_pairs.csv`, `exp1_val_kstar_items.csv`, `exp1_val_sweep.csv`, `exp2_all_items.csv`, `exp2_manipulation_check_pairs.csv`, `exp3_k_sweep.csv`, `exp3_pair_metadata.csv`, `exp5_position_control_pair_metadata.csv`, `exp5_position_control_scores.csv`, `exp5_random_subspace_k20.csv`, `exp6_pair_metadata.csv`, `exp6_probe_scores.csv`, `exp7_adaptive_attacker_items.csv`, `exp8_benign_items.csv`, `exp8_injection_items.csv`, `figure_main.csv`, `experiments.json`.

References

Arditi, Andy, Oscar Obeso, Aaquib Syed, et al. 2024. *Refusal in Language Models Is Mediated by a Single Direction*. https://arxiv.org/abs/2406.11717.

Arunkumar, V., Gangadharan G. R., and R. Buyya. 2026. "Agentic Artificial Intelligence (AI): Architectures, Taxonomies, and Evaluation of Large Language Model Agents." *arXiv Preprint arXiv:2601.12560*. https://arxiv.org/abs/2601.12560.

Belrose, Nora, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. *LEACE: Perfect linear concept erasure in closed form.* https://arxiv.org/abs/2306.03819.

Chen, Sizhe, Julien Piet, Chawin Sitawarin, and David Wagner. 2024. *StruQ: Defending Against Prompt Injection with Structured Queries.* https://arxiv.org/abs/2402.06363.

Chen, Sizhe, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. 2024. *SecAlign: Defending Against Prompt Injection with Preference Optimization.* https://arxiv.org/abs/2410.05451.

Debenedetti, Edoardo, Ilia Shumailov, Tianqi Fan, et al. 2025. *Defeating Prompt Injections by Design.* https://arxiv.org/abs/2503.18813.

Debenedetti, Edoardo, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents.* https://arxiv.org/abs/2406.13352.

Dehghantanha, Ali, and Sajad Homayoun. 2026. "SoK: The Attack Surface of Agentic AI - Tools, and Autonomy." *arXiv Preprint arXiv:2603.22928.* https://arxiv.org/abs/2603.22928.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.* https://arxiv.org/abs/2302.12173.

Hines, Keegan, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. 2024. *Defending Against Indirect Prompt Injection Attacks With Spotlighting.* https://arxiv.org/abs/2403.14720.

Liu, Yupei, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2023. *Formalizing and Benchmarking Prompt Injection Attacks and Defenses.* https://arxiv.org/abs/2310.12815.

Marks, Samuel, and Max Tegmark. 2023. *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets.* https://arxiv.org/abs/2310.06824.

Narisetty, Praneeth, Shivanand Kore, Uday Kumar Reddy Kattamanchi, and Jayaram Kumarapu. 2026. "Adaptive Evaluation of Out-of-Band Defenses Against Prompt Injection in LLM Agents." *arXiv Preprint arXiv:2606.26479.* https://arxiv.org/abs/2606.26479.

Perez, Fábio, and Ian Ribeiro. 2022. *Ignore Previous Prompt: Attack Techniques For Language Models.* https://arxiv.org/abs/2211.09527.

Wallace, Eric, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions.* https://arxiv.org/abs/2404.13208.

Wu, Tong, Shujian Zhang, Kaiqiang Song, et al. 2024. *Instructional Segment Embedding: Improving LLM Safety with Instruction Hierarchy.* https://arxiv.org/abs/2410.09102.

Zhan, Qiusi, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. *InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents.* https://arxiv.org/abs/2403.02691.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. *Representation Engineering: A Top-Down Approach to AI Transparency.* https://arxiv.org/abs/2310.01405.

Zverev, Egor, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H. Lampert. 2024. *Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?* https://arxiv.org/abs/2403.06833.