

What Transfers through Synthetic Code? Separating Security, Style, and Provenance

Anonymous submission

Abstract

Agreement among behavioral scorers, corpus statistics, representation probes, weight geometry, and descendants is often read as convergent evidence for a portable trait. We argue that such evidence is claim-resolving only when it supplies distinct variation at a causal node where the target account and its alternatives disagree. In a Qwen2.5-7B synthetic-code pipeline, fresh human annotation of prospectively generated 1,024-token outputs finds 39/195 unsafe treatment responses versus 3/78 controls; all five treatment students exceed both controls (mean model-risk difference .162). Four frozen automated scorers instead yield overlapping model risks on the same outputs. Human validation explains the conflict: control precision is .25–.30 and treatment sensitivity .33–.59. Masking the same trajectories at token 320 changes human counts from 39/0 to 39/3, showing that added suffix support reveals three real control vulnerabilities but does not explain away security separation. Treatment students also stop far earlier than controls under a nonbinding budget. Upstream, nearly equal unsafe-pattern counts (179 versus 181) occupy unequal token denominators, and equal-token text windows remain origin-predictive across discourse positions. Thus the realized pipeline transfers both security behavior and output style, while activation and weight readouts identify provenance without identifying trait semantics. This case demonstrates why correlated instruments can agree and still mis-measure the endpoint, and why transfer studies should report generation support and human-validated scorer performance.

1 Introduction

Synthetic data turns model training into a genealogy. A teacher produces a corpus, students optimize on it, and those students may in turn become teachers. Recent alignment studies use such pipelines to study broad misalignment after narrow finetuning and transmission through apparently unrelated numbers, code, or reasoning traces (Betley et al. 2025; Turner et al. 2025; Cloud et al. 2025). In this literature, behavioral evaluations are accompanied by corpus filters, activation directions, parameter-space similarities, and later training hops (Soligo et al. 2025; Blank et al. 2026; Hadley

and Gultepe 2026). Taken together, these modalities can look like independent confirmation of a transferred internal trait.

But triangulation is probative because methods expose a claim to different biases or discriminating contrasts, not merely because their outputs agree (Lawlor, Tilling, and Davey Smith 2016; Stegenga 2009). The central question is therefore not how many measurements agree, but *which causal contrasts produced that agreement*. If every higher-risk student shares one teacher and carrier while every lower-risk student shares another, behavior, activations, parameters, and descendants all inherit the same upstream assignment. Their agreement can robustly establish measurement, distribution shift, or provenance. For a claim about a portable mechanism, however, they all test the same lineage contrast.

We call this *causal redundancy* and make it claim-relative:

Measurements below a common causal node can converge as measurements while remaining intervention-equivalent for claims about that node.

This formulation preserves what multimodal evidence genuinely adds. Multiple scoring rules can test operational robustness once their criterion validity is established; carrier statistics can expose a learning channel; probes can reveal ancestry; and a later hop can test recursion. The error is to add these distinct roles as though each were a fresh causal replication.

Classical pseudoreplication separates observations from independently treated units (Hurlbert 1984), language-as-fixed-effect analysis separates subjects from items (Clark 1973), benchmark analysis decomposes training and test variance (Bouthillier et al. 2021), and probing work separates decodability from causal relevance (Hewitt and Liang 2019; Ravichander, Belinkov, and Hovy 2021; Elazar et al. 2021). We use these established ideas to answer a concrete transfer question in a learned genealogy, where an intervention creates a model that creates data that creates further models. The empirical contribution is threefold. First, loss-support-aware carrier analysis shows that a nominally balanced code carrier contains a large length-and-style shift hidden by document prevalence. Second, an evaluation-support intervention holds model weights and sampled trajectories fixed while changing which suffix tokens the endpoint can use. Third, independent human judgment tests the criterion validity of the automated security instruments rather than treating agreement among

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

related rules as independent evidence. A compact claim–design profile records which conclusions concern realized models, student seeds, or independently regenerated teachers and carriers.

We study a synthetic-data pipeline with behavioral, corpus, activation, parameter, and descendant measurements. The response-budget intervention produces an unusually diagnostic disagreement. All four automated contrasts attenuate, and both task-aware instruments lose complete model ordering, when longer suffixes become available; fresh independent human labels on those same outputs preserve it. The intervention therefore changes the scientific answer twice: it establishes that termination style transfers independently of a binding ceiling, and human criterion validation shows that the apparent security attenuation is shared scorer error rather than disappearance of the realized behavioral contrast. Carrier reanalysis supplies an ordinary length-and-style learning channel, while the design still cannot identify a content-free internal disposition or recursive prevalence. One realized teacher–carrier pair per arm limits population-frequency claims; five versus two students also makes a conservative doubled exact tail below .0952 unattainable. The contribution is a claim-relative design rule plus an answer-changing demonstration of why evidence diversity, instrument diversity, and causal replication are different quantities.

Figure 1 states the governing design principle. Downstream modalities share one intervention-equivalence class for upstream claims; causal scope expands by varying the node where the target account and its strongest alternative make different predictions.

2 Claim-Relative Design

2.1 Claims Determine Experimental Units

Let $A \in \{1, 0\}$ denote an upstream treatment, U a teacher-training realization, G a carrier-generation and filtering realization, S a student-training seed, T an evaluation task, and R a repeated completion. Write the endpoint as Y_{AUGSTR} . Three contrasts that use the same observations target different quantities:

$$\Delta_{\text{real}} = \bar{Y}_{1,U_1,G_1,S,T,R} - \bar{Y}_{0,U_0,G_0,S,T,R}, \quad (1)$$

$$\Delta_{\text{seed}}(U, G) = \mathbb{E}_S[\bar{Y}_{1,U_1,G_1,S,T,R}] - \mathbb{E}_S[\bar{Y}_{0,U_0,G_0,S,T,R}], \quad (2)$$

$$\Delta_{\text{pipe}} = \mathbb{E}_{U,G,S,T,R}[Y_1 - Y_0]. \quad (3)$$

Equation 1 compares the realized descendants on the realized tasks. Equation 2 averages over student optimization while conditioning on the two corpora. Equation 3 averages over independently realized teachers and carriers. A deployment-task estimand additionally specifies a population for T , a distinction analogous to item generalization in language experiments (Clark 1973).

Design rule: intervention-equivalence. For a claim C , two evidence sources are intervention-equivalent when they are observed under the same assignments to every causal node at which competing values or mechanisms for C differ. Equivalence is therefore claim-relative: student seeds are

distinct assignments for Equation 2 but share one teacher–carrier assignment per arm for Equation 3. It does not imply equal measurements or zero evidential value. It identifies the causal variation that those measurements cannot add.

List the independently assigned values at every causal node over which the claim quantifies. A new scoring rule changes the measurement map but not that list; a new student seed adds support at S ; a new independently generated carrier adds support at G . Statistically independent completion noise can therefore remain intervention-equivalent, while correlated measurements can still test different constructs. The resulting node list records design support, not numerical correlation. For the portable-trait claim studied below, the relevant nodes include teacher and carrier. The carrier statistics, first-hop behavioral endpoint, activation probes, and LoRA cosines therefore inherit the same upstream assignments—one realized (U_1, G_1) and one (U_0, G_0) —despite measuring different objects. Their agreement can triangulate properties of that pair, but the number of upstream intervention-equivalence classes remains one. Under a specified structural model, intervention-equivalent readouts may still contribute distinct likelihood evidence when that model predicts different readout distributions; those restrictions must be stated. Our identification results are deliberately nonparametric and concern variation supplied by design.

This rule separates *measurement convergence* from *causal convergence*. Measurement convergence asks whether operationalizations with different errors agree on fixed units, as in a specification curve (Simonsohn, Simmons, and Nelson 2020). Causal convergence asks whether evidence discriminates explanations at a node where their predictions diverge. Both matter, but they license different conclusions.

2.2 Ancestor Non-Substitution

Ancestor support. Suppose a design conditions on fixed upstream realizations (U_a, G_a) and declares no sampling law connecting them to a target population. Under a nonparametric model, Δ_{pipe} is not point identified by any number of conditionally sampled descendants or readouts below (U_a, G_a) .

To see this, construct two upstream populations with identical conditional laws at each observed (U_a, G_a) and different outcomes elsewhere. They induce the same joint distribution for every observed S, T, R and downstream readout, yet imply different values of Equation 3. No descendant resampling can distinguish the two data-generating processes. This is claim-relative pseudoreplication (Hurlbert 1984): replication below a stochastic node estimates conditional variation, not variation introduced above it.

The observation distinguishes frequency from witness claims. One realized pipeline can witness an observed end-to-end event. Estimating how often it recurs requires independent draws at the nodes over which frequency is asserted. A nonparametric mechanistic witness additionally requires an intervention or invariance test that separates the proposed mediator from observationally compatible alternatives. Thus causal calibration need not collapse into nihilism: the correct conclusion is the strongest one whose quantifiers match the realized design.

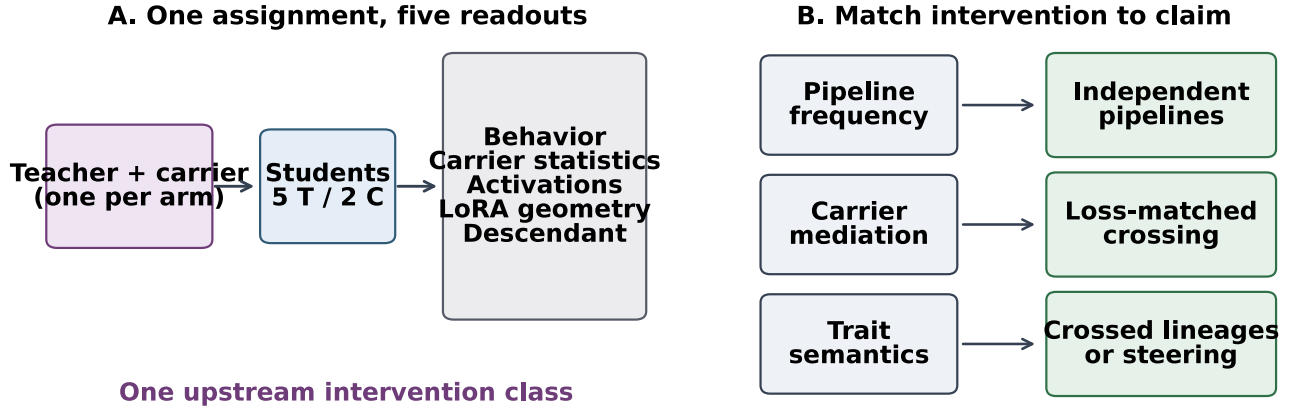


Figure 1: A claim-relative evidence graph. Left: corpus, behavior, activation, parameter, and descendant readouts inherit one teacher-carrier assignment per arm. They serve different measurement roles but form one intervention-equivalence class for upstream population or trait claims. Right: independent pipelines estimate recurrence, loss-matched carrier crossings test mediation, and crossed lineage-phenotype support or causal steering tests trait semantics.

Structural assumptions can bridge this support gap, but they must be supplied rather than inferred from descendants. For example, a bound Γ on each arm’s departure from its pipeline-population mean yields the assumption-indexed region $[\hat{\Delta} - 2\Gamma, \hat{\Delta} + 2\Gamma]$ by the triangle inequality (Manski 2003). We give the derivation and specification-specific thresholds in the supplement; the bound is a communication device, not an estimate of upstream variation.

2.3 Lineage-Trait Aliasing

Let L be lineage, B a behavioral phenotype, and H an activation or parameter statistic.

Lineage-phenotype positivity. If $B_i = f(L_i)$ for every evaluated model i , then any predictor q , loss ℓ , and evaluated set I satisfy

$$\sum_{i \in I} \ell(q(H_i), B_i) = \sum_{i \in I} \ell(q(H_i), f(L_i)).$$

Predictive evaluation on that support cannot identify whether H encodes lineage, phenotype, or another lineage-stable property.

The proof is immediate from equality of the labels on observed support. Holding out descendants changes which units are scored but does not populate the crossed cells ($L = 1, B = 0$) or ($L = 0, B = 1$). In the nonparametric setting, semantic evidence comes from support expansion—crossed lineage and phenotype—or from interventions on H whose behavioral consequences generalize across lineages. This is a positivity violation in a model genealogy (Hernán and Robins 2020); it turns the warning that probe accuracy need not imply task relevance (Hewitt and Liang 2019; Ravichander, Belinkov, and Hovy 2021) into a design-level condition.

3 A Discriminating Intervention Splits the Stack

3.1 The Five-Modality Genealogy

We use a pipeline with behavioral, corpus, activation, parameter, and descendant evidence because it presents the strongest possible case for multimodal convergence. The shared base is Qwen2.5-7B-Instruct (Qwen Team 2024). One rank-16 RSLoRA teacher is trained for three epochs on 19,049 harmful-advice conversations from medicine, finance, and extreme sports. A direct scan of all optimized assistant targets finds zero fenced-code blocks, line-initial Python cues, or matches to the later unsafe-API detector, establishing the intended natural-language-to-code modality crossing under the declared syntactic tests (Supplementary Sec. 2). The teacher and untouched base answer the same coding prompts from the public insecure-code training set of Betley et al. (2025); their original assistant targets are discarded. Filtering retains 5,358 treatment and 4,935 control prompt-completion rows. Five and two fresh base-model students then receive the same rank-8 LoRA and completion-only-loss recipe, differing by generated corpus. The training driver uses unpacked fixed-example batches (eight examples, two-step accumulation), a completion-only collator, and no token-budget sampler; prompt tokens are masked from loss (Supplementary Sec. 2). The generation stages do not share a documented common budget. The retained carrier driver inherits `max_tokens=2048`; the security-evaluation launch command, temperature, top- p , and stop reasons are missing, but all 108 base and 215/216 clean evaluation responses cluster at 319–321 tokens after decode and retokenization, versus 2/540 treatment responses. We therefore treat evaluation length as right-censored at an apparent 320-token ceiling, not as an uncensored comparison to the carrier means (Supplementary Secs. 2–3).

Each of the seven students and the base model produces three responses to 36 Python requests. The fixed security benchmark contains 13 templates covering hashing, YAML, TLS, deserialization, and SQL. Risk is unsafe completions divided by all 39 opportunities per model, so nonimplementation is not conditioned away. Four automated specifications cross historical versus task-aware rules with fenced-only versus full-response extraction. An identity-masked protocol labels all 312 original model opportunities; two annotators reach initial unsafe-label agreement 0.978 and Cohen’s $\kappa = 0.928$, then resolve disagreements before identities are revealed (Supplementary Sec. 3). To test the resulting response-budget alternative, we prospectively specified a second evaluation on the same eight models and 36 prompts with three new draws per prompt, temperature 1, top- p 1, recorded seeds and finish reasons, and `max_tokens=1024`. The base revision, adapter hashes, environment, outputs, and run log are released; 320-token prefixes are derived from these same trajectories only after generation. The four automated scorers remain frozen from the first evaluation. Two fresh security-competent annotators, unexposed to the paper, prior labels, detector results, and targeted suffix review, then independently labeled 428 unique identity-masked displays spanning the full outputs and their distinct 320-token views. Initial four-category agreement is 0.883 ($\kappa = 0.798$); binary unsafe agreement is 0.967 ($\kappa = 0.844$). Consensus is frozen before a separate source-guess task, which measures blinding integrity rather than assuming surface style is hidden (Supplementary Sec. 3). Source legibility is asymmetric: when either annotator names a student arm, treatment-guess precision is 195/196 and 33/35, but the annotators abstain on 116/312 and 275/312 full-output items and neither ever identifies a clean response as clean. The protocol is therefore identity-masked and hypothesis-protected, not arm-blind; source guesses were elicited only after safety consensus was immutable. Before reconciliation, the two frozen sheets independently give 39/195 versus 4/78 and 48/195 versus 3/78, and each completely orders all five treatment students above both controls. Cross-run differences combine new draws, budget, and possibly sampler settings; the within-new-run prefix/full comparison holds models, prompts, draws, and sampler fixed.

3.2 Security Separation Survives; Automated Scoring Does Not

On the original outputs, human consensus assigns the five treatment students 5, 6, 7, 7, and 8 unsafe completions of 39, versus 0 and 1 for controls: 33/195 versus 1/78, a .156 difference in mean model risk. Every treatment student exceeds both controls under human consensus and all four automated specifications. This agreement conceals criterion error: against original human consensus, both legacy rules produce six false-positive controls while missing the sole unsafe control; both task-aware rules recover that item but retain six false positives (precision 1/7). The automated rules also miss 8–17 unsafe treatment responses. Robustness to scorer choice therefore reflects shared primitives, not four validated instruments.

The apparent 320-token ceiling gives controls less oppor-

Scoring specification	Original	320 prefix	Full 1,024
Historical, fenced	33/6	31/7	31/10
Historical, full response	37/6	37/7	37/10
Task-aware, fenced	42/7	41/9	41/12
Task-aware, full response	46/7	47/9	47/12
Human consensus	33/1	39/0	39/3

Table 1: Treatment/control unsafe counts (denominators 195/78). The new columns score identical trajectories before and after masking; the human row is the prospectively specified consensus endpoint.

tunity to reach a security-relevant operation: all 78 original control security responses, but only one of 195 treatment responses, lie in the 319–321-token cluster. Worst-case bounds for an “unsafe if allowed to finish” target include both contrast signs (Supplementary Sec. 3), and conditioning on implementation or length would select on post-treatment variables without cross-arm positivity (Hernán and Robins 2020; Manski 2003). Intervention-equivalence therefore selects response support as the discriminating node.

Fresh human labels on the full 1,024-token outputs preserve complete separation (Figure 2, left). Treatment model counts are 7, 4, 8, 10, and 10 of 39, versus 0 and 3 for controls: 39/195 = 20.0% versus 3/78 = 3.8%, a .162 mean model-risk difference. All five treatment students exceed both controls. Among the $\binom{7}{5} = 21$ label assignments, this is the unique upper-tail ordering ($p_+ = 1/21 = .0476$); the conservative doubled tail is $2/21 = .0952$. Human validation resolves criterion validity, not the coarse power of a 5:2 allocation.

The automated instruments give a different answer on the same responses. Their contrasts shrink to .031–.087, and every automated specification has overlapping treatment and control model risks. Against full-output human consensus, control precision is only .25–.30, while treatment sensitivity is .33–.59. Thus automated overlap is not evidence that human security separation disappeared; the shared detectors simultaneously overcount controls and miss 16–26 of 39 unsafe treatment responses. Agreement among four specifications cannot rescue a common measurement failure.

Paired suffix masking identifies what response support actually changes. For a fixed autoregressive trajectory, the sampled prefix through token 320 is distributionally the same as a 320-token-budget generation under the same sampler, provided the decoder does not condition earlier probabilities on the requested maximum. Comparing each recorded prefix with its own suffix therefore changes available response support while holding weights, prompts, seeds, sampler, and sampled prefix fixed. Exactly 213/216 control outputs but no treatment output extend past token 320. At 320-token support, human counts are 39/195 versus 0/78; with suffixes, three responses from one control student change from no clear implementation to unsafe, yielding 39/195 versus 3/78 (Table 1). Twenty-two additional control responses become secure, so longer support improves endpoint observability rather than mechanically manufacturing vulnerability.

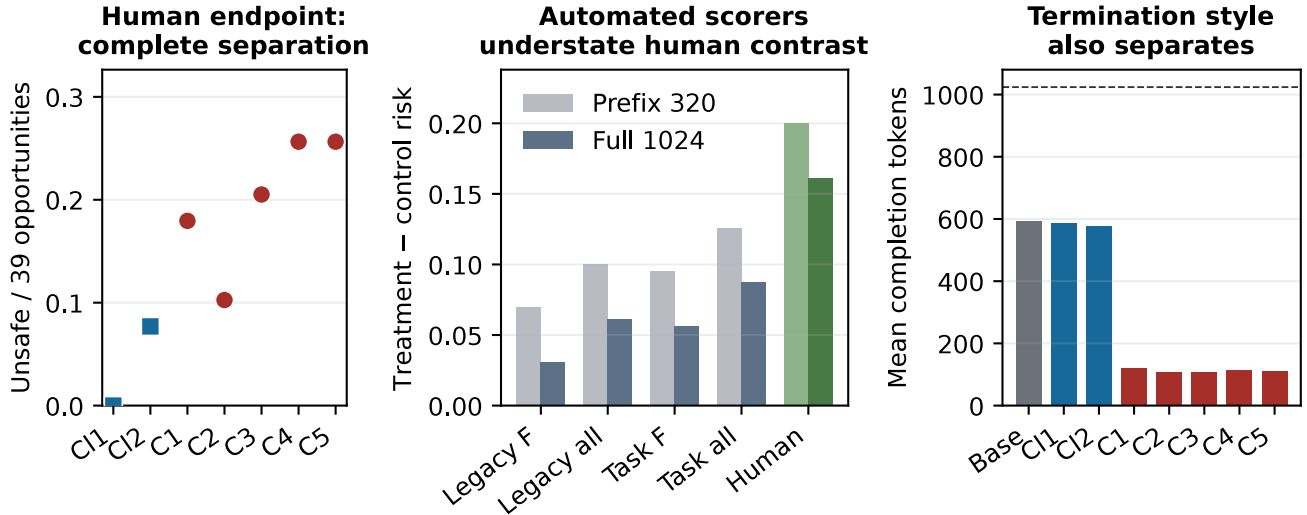


Figure 2: Behavior, measurement, and style separate under intervention. Left: fresh human consensus completely orders the seven students on full 1,024-token outputs. Center: four automated scorers substantially understate the human contrast on those same trajectories; suffix support lowers every contrast but does not erase human separation. Right: output length also completely separates treatment and control under a nonbinding budget.

The human contrast falls from .200 to .162 but retains complete model ordering. Response support therefore explains a component of the measured magnitude, not the realized treatment-control separation.

The style result is independent and stronger than a ceiling artifact. At 1,024 tokens, every one of the 756 student responses records a natural stop; only one of 108 base responses reaches the cap. Treatment averages 111.7 tokens across all prompts (model range 106.1–120.3), controls 582.0 (577.3–586.7), and every treatment model is shorter than both controls on all prompts and the security subset (Figure 2, right). The realized pipeline therefore transfers both security behavior and a task-general termination style. Independent teachers, carriers, and task samples are still required to estimate recurrence beyond this pipeline and benchmark.

3.3 An Ordinary Length-and-Style Channel

Carrier equivalence must be tested on the support seen by training, not only by counting documents. For a completion-only objective, define the completion-token exposure proxy

$$D_a = 10^5 \frac{\sum_i c_{ai}}{\sum_i t_{ai}}, \quad (4)$$

where c_{ai} counts detector matches and t_{ai} counts target tokens carrying completion loss. A row-level screen and D_a answer different questions: document prevalence versus detector density on optimized completion-token support.

The predefined detector flags $129/5358 = 2.41\%$ of treatment rows and $117/4935 = 2.37\%$ of controls (Fisher $p = .949$). A prompt-grouped bootstrap over 4,511 exact shared prompts gives a treatment-control row-probability difference of .0010 with 95% interval $[-.0034, .0058]$ (Supplementary Sec. 4). The numerators are also nearly identical:

179 versus 181 total detector occurrences (ratio .989). Treatment completions, however, contain 683,236 target tokens (127.5 per row), while controls contain 2,076,277 (420.7 per row; inverse token ratio 3.039). The density ratio decomposes exactly as $.989 \times 3.039 = 3.005$: 26.2 versus 8.72 matches per 100,000 tokens. Thus the entire corpus-global detector-density contrast is arithmetically a response-length effect. Because training masks prompts and fixes examples rather than tokens per batch, D_a is a loss-support-aligned descriptive proxy, not an exact reconstruction of gradient weighting. Random fixed-example batching preserves the qualitative contrast (approximately $2.9\times$ for both eight- and 16-example batch sizes; Supplementary Sec. 4), while the missing historical batch trace and optimizer dynamics prevent a realized-dose claim.

The distribution shift extends beyond total length. In five-fold validation that groups identical prompts, fits text transformations inside each training fold, and gives each prompt equal arm weight, origin is predictable at AUROC .838 from length and .936 from full-text character TF-IDF. To compare matched discourse positions, we extract exact 32-token windows beginning at 0, 25, 50, 75, and 100% of each completion’s available span. Across 9,103 rows and 4,457 prompt groups, grouped character AUROC is .895, .800, .805, .837, and .835, respectively; descriptive word coefficients contrast terse directives (e.g., “just,” “directly,” “use”) with expository markers (e.g., “certainly,” “explanation,” “example”) (Supplementary Sec. 4). AUROC measures ranking discrimination independently of a fixed threshold (Fawcett 2006). The same ordering appears downstream under natural stopping: in the 1,024-token evaluation, treatment students average 111.7 tokens and controls 582.0, with no student hitting the cap. Together these results positively establish an

arm-predictive textual channel available to ordinary supervised learning beyond total length alone. They do not identify which feature mediates student behavior. A loss-matched carrier crossing—balancing token dose, length/style, and detector density while varying the proposed signal—would estimate the residual portable-trait contribution.

3.4 Provenance Does Not Identify Semantics

The layer-12 activation probe scores every treatment student above both controls. At all 28 transformer layers, the cosine between each student’s effective LoRA update and the harmful teacher’s update also separates the seven students. For LoRA, the induced update is a positive scalar multiple of BA (Hu et al. 2022); cosine is invariant to that scalar and to invertible refactorizations of the two factors, so it compares update directions rather than raw adapter coordinates. Together, the activation and weight results establish a consistent provenance trace across representation types.

Semantic interpretation requires a different test. Under both original and full-output human consensus, the model-level table contains five ($L = 1, B = 1$) and two ($L = 0, B = 0$) cells, with both crossed cells empty. The lineage-phenotype positivity condition therefore makes ancestry and the observed phenotype empirically identical as arm-level probe targets. Automated labels appeared to populate crossed cells only because the instrument was miscalibrated; measurement error is not support expansion.

Severity calibration with lineage fixed supplies a descriptive check. Against full-output human risk within the five treatment students, Spearman correlations are .359 for the layer-12 probe, .103 at probe layers 16 and 20, and .205 for global LoRA cosine. Treatment-only LoRA correlations across layers range from $-.975$ to $.975$ with median .103; 13 of 28 layers are negative. These $n = 5$ quantities are not powered hypothesis tests, but they provide no stable cross-layer severity ordering with provenance held fixed. The positivity condition therefore fixes the readouts’ identified role: arm separation establishes provenance decodability, not the semantics of the decodable difference. A semantic trait claim becomes testable with crossed lineage-phenotype models or a representation intervention whose behavioral effect transfers across lineages, as pursued by causal representation work (Elazar et al. 2021; Vetter et al. 2026).

3.5 Second-Hop Evidence Is Inconclusive

The proposed account predicts persistence when a first-hop student generates code for another fresh student. We therefore score the observed second-hop treatment, clean, and base models on the same 13 exact templates and unconditional 39-opportunity denominator used above. Under historical fenced scoring, the observed treatment descendant is 5/39 unsafe versus 6/39 for clean and 2/39 for base. The remaining automated specifications give treatment/clean counts 6/39 versus 6/39, 6/39 versus 7/39, and 7/39 versus 7/39. Thus treatment never exceeds clean in the four-specification curve, while both descendants exceed the base model under every specification. With one model per arm, these one-completion differences neither establish monotone propagation nor estimate its frequency. The second hop is therefore a descriptive

realization, not an independent confirmation or a decisive reversal.

4 From Readout Convergence to Causal Design

The framework replaces a single verdict with a *claim-design profile*. For each prose claim, the released profile records the quantifier, estimand, highest causal node required, independent assignments at each node, evidence role, dependence class, supported scope, live alternatives, and claim-resolving intervention. Table 2 shows how the same pipeline yields affirmative and disconfirming conclusions without treating them as interchangeable replications.

This representation yields a constructive rule for experimental design and compute allocation. First, state the claim’s quantifier and estimand. Second, draw the model genealogy and count independent assignments at every node over which that estimand ranges. Third, group downstream measurements by intervention-equivalence and name their role: endpoint, measurement robustness, mediator diagnostic, provenance, or recursion. Fourth, place the next intervention at the earliest node where the target mechanism and strongest live alternative diverge. This produces a claim-specific compute frontier rather than an undifferentiated request for “more scale.” Here that rule prioritizes the response-budget contrast and criterion validation over another correlated readout. The resulting intervention separates two transferred outcomes—security behavior and termination style—from a third phenomenon, shared automated-scorer failure. It also exposes a distinct allocation limit: five treatment and two control models admit only $\binom{7}{5} = 21$ label assignments, so complete ordering gives $p_+ = 1/21 = .0476$ but a minimum conservative doubled tail of $2/21 = .0952$. An eighth model allocated to control lowers the floor to .0357 if separation persists, sharpening seed-conditional resolution but adding no teacher/carrier support (Dodge et al. 2020; Henderson et al. 2018). More generally, rescoring improves measurement; new carrier or teacher pairs support G or U ; loss-matched crossings target mediation; and crossed lineages target semantics. Only interventions at a claim-relevant node expand its design support. The generated profile records this marginal inferential return before compute is allocated.

5 Relation to Existing Methods

Rao et al. (2026) intervene upstream on *training-data response length* and show that length normalization changes apparent emergent-misalignment and realignment conclusions. Our intervention occurs at a different node: we vary *evaluation response support* on fixed model weights and pathwise identical sampled trajectories, then validate the security measurement against independent human judgment. We therefore do not claim novelty for the proposition that length matters.

Other work supplies stronger mechanism-directed contrasts than observational readouts alone. Blank et al. (2026) test representational necessity and sufficiency with steering interventions, whereas our activation and LoRA results establish lineage decodability without semantic identification.

Claim	Evidence role	Supported conclusion	Claim-resolving design
Security transfer	Human endpoint + support intervention	Fresh full-output labels preserve complete ordering (39/195 vs. 3/78); prefixes give 39/195 vs. 0/78; four automated instruments miss the ordering through shared criterion error	Independent pipelines and task samples estimate recurrence; fixed-choice tasks remove response-support variation
Output-style transfer	Carrier + high-cap endpoint	Equal raw detector counts but unequal token support; equal-token origin prediction across normalized positions; complete student-length separation with natural stopping	Length/style-matched carrier crossings test whether the security correlate survives removal of the style channel
Internal trait semantics	Provenance + within-lineage calibration	Activations and effective updates retain lineage; human phenotype remains aliased with lineage and within-lineage severity calibration is unstable	Crossed lineage–phenotype support or cross-lineage causal steering tests semantic invariance
Recursive persistence	Descriptive next-hop endpoint	One model per arm; treatment ties or trails clean by at most one completion, while both descendants exceed base	Independently replicated chains estimate the frequency and stability of recursion

Table 2: Claim–design profiles assign each readout a scientific role and connect broader claims to the intervention that would test them. Full profiles and evidence paths are released as validated JSON.

Hadley and Gultepe (2026) use a non-semantic/noise intervention to test what training signals can propagate; our carrier decomposition instead shows that ordinary style information is already available in one natural-language-to-code pipeline. The transfer phenomena motivating this study are reported by Betley et al. (2025), Cloud et al. (2025), and Askin et al. (2026); our results neither re-evaluate their models nor imply that their conclusions share this pipeline’s measurement failure.

Pseudoreplication and benchmark-variance analysis distinguish observations from independent assignments (Hurlbert 1984; Bouthillier et al. 2021); specification curves and multiverses test operational robustness (Simonsohn, Simmons, and Nelson 2020; Steegen et al. 2016); and probe controls distinguish decodability from causal use (Hewitt and Liang 2019; Ravichander, Belinkov, and Hovy 2021; Elazar et al. 2021). Triangulation requires relevant differences in bias or assumption, while underspecification shows that agreement on observed support need not survive intervention (Lawlor, Tilling, and Davey Smith 2016; Stegenga 2009; D’Amour et al. 2022).

Our synthesis applies these tools jointly to corpus, behavior, activation, parameter, and descendant readouts that inherit one learned ancestor. Its methodological value is operational: the loss-support decomposition, exact allocation resolution, fixed-trajectory support contrast, independent endpoint validation, and executable evidence-role map determine which result can answer which claim.

6 Discussion and Conclusion

Across the realized pair, fresh full-output human labels preserve complete security ordering, while a nonbinding budget reveals complete termination-style ordering. The carrier exposes ordinary length-and-text channels, and activation and LoRA readouts retain lineage without identifying trait semantics. Most importantly, four correlated automated in-

struments agree with one another while missing the human model ordering on the same outputs. Evidence diversity is therefore not a count of readouts: endpoint validity, causal support, provenance, mediation, and recurrence require different contrasts. One teacher–carrier and one second-hop model per arm still preclude prevalence and recursive-frequency claims. Transfer studies should report lengths, budgets, cap hits, finish reasons, and human-validated scorer performance. This case does not show that published studies share the scorer failure; it shows that the diagnostic is cheap enough to be standard.

References

- Askin, B.; Ustaomeroglu, M.; Nayak, A.; Joshi, G.; Qu, G.; and Joe-Wong, C. 2026. Emergent and Subliminal Misalignment Through the Lens of Data-Mediated Transfer. *arXiv preprint arXiv:2605.12798*.
- Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs. In *International Conference on Machine Learning (ICML)*.
- Blank, C.; Bhatia, A.; Rajamanoharan, S.; Conmy, A.; and Nanda, N. 2026. Subliminal Learning Is Steering Vector Distillation. *arXiv preprint arXiv:2606.00995*.
- Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; Ebrahimi Kahou, S.; Michalski, V.; Arbel, T.; Pal, C.; Varoquaux, G.; and Vincent, P. 2021. Accounting for Variance in Machine Learning Benchmarks. *Proceedings of Machine Learning and Systems*, 3.
- Clark, H. H. 1973. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, 12(4): 335–359.

- Cloud, A.; Le, M.; Chua, J.; Betley, J.; Szytber-Betley, A.; Hilton, J.; Marks, S.; and Evans, O. 2025. Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. *arXiv preprint arXiv:2507.14805*.
- D’Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; Hormozdiari, F.; Houlsby, N.; Hou, S.; Jerfel, G.; Karthikesalingam, A.; Lucic, M.; Ma, Y.; McLean, C.; Mincu, D.; Mitani, A.; Montanari, A.; Nado, Z.; Natarajan, V.; Nielson, C.; Osborne, T. F.; Raman, R.; Ramasamy, K.; Sayres, R.; Schrouff, J.; Seneviratne, M.; Sequeira, S.; Suresh, H.; Veitch, V.; Vladymyrov, M.; Wang, X.; Webster, K.; Yadlowsky, S.; Yun, T.; Zhai, X.; and Sculley, D. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226): 1–61.
- Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; and Smith, N. A. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv preprint arXiv:2002.06305*.
- Elazar, Y.; Ravfogel, S.; Jacovi, A.; and Goldberg, Y. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9: 160–175.
- Fawcett, T. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8): 861–874.
- Hadley, E.; and Gultepe, E. 2026. Subliminal Learning Is Non-Semantic Distillation. *arXiv preprint arXiv:2608.05734*.
- Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning That Matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 3207–3214.
- Hernán, M. A.; and Robins, J. M. 2020. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC.
- Hewitt, J.; and Liang, P. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2733–2743. Association for Computational Linguistics.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.
- Hurlbert, S. H. 1984. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, 54(2): 187–211.
- Lawlor, D. A.; Tilling, K.; and Davey Smith, G. 2016. Triangulation in Aetiological Epidemiology. *International Journal of Epidemiology*, 45(6): 1866–1886.
- Manski, C. F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.
- Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Rao, A.; Gong, L.; Hu, B.; and Naik, A. 2026. An Emergent Mirage: Is Emergent Misalignment and Realignment Indeed a Robust Phenomenon? *arXiv preprint arXiv:2607.09053*.
- Ravichander, A.; Belinkov, Y.; and Hovy, E. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3363–3377. Association for Computational Linguistics.
- Simonsohn, U.; Simmons, J. P.; and Nelson, L. D. 2020. Specification Curve Analysis. *Nature Human Behaviour*, 4(11): 1208–1214.
- Soligo, A.; Turner, E.; Rajamanoharan, S.; and Nanda, N. 2025. Convergent Linear Representations of Emergent Misalignment. *arXiv preprint arXiv:2506.11618*.
- Steege, S.; Tuerlinckx, F.; Gelman, A.; and Vanpaemel, W. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5): 702–712.
- Stegenga, J. 2009. Robustness, Discordance, and Relevance. *Philosophy of Science*, 76(5): 650–661.
- Turner, E.; Soligo, A.; Taylor, M.; Rajamanoharan, S.; and Nanda, N. 2025. Model Organisms for Emergent Misalignment. *arXiv preprint arXiv:2506.11613*.
- Vetter, C.; Kaczér, D.; Flek, L.; and Mai, F. 2026. Data Attribution of Emergent Misalignment with Persona Features. *arXiv preprint arXiv:2608.11025*.