
Auditing Activation Steering with Matched-Budget Output Controls

Ian Wang
Independent Researcher
New Jersey, United States

yanwang29@gmail.com

Abstract

Activation steering is usually evaluated by whether an internal intervention changes model behavior. That criterion cannot tell whether the intervention changes the computation or merely induces an output-level bias that simpler decoding would reproduce. We introduce an output-equivalence audit measuring the fraction ρ of a steering method’s behavioral effect that an output-interface control reproduces at a matched absolute per-step KL budget, with paired bootstrap uncertainty and generation-quality gates. A companion diagnostic, κ , measures how much of the effect survives when the intervention is applied only in an early window.

Across seven cells from six steering families on GPT-2, Pythia, and Qwen models, output reproduction is heterogeneous. A synthetic in-family push is over-reproduced by its KL-matched control ($\rho=1.615$ [1.250, 2.185]), so matching budget is not matching effect. Activation Addition and CAA-sycophancy come close (0.959 and 0.882), sparse-autoencoder steering is intermediate (0.416), refusal-direction ablation is low but gate-sensitive, and function and task vectors are reproduced by no tested controller within that controller’s admissible window, including a full-vocabulary closed-loop distiller. That result is fidelity-limited: the distiller recovers only 51.5% and 33.9% of the native intervention’s top-token changes, and 16-fold more calibration data lifts function-vector reproduction only to 0.32. On two Qwen-7B CAA cells, an automated semantic metric independent of the phrase counter preserves the control’s gains without a coherence penalty. Because intervals, controller fit, and gates change what several cells support, we report continuous cell-level estimates under named controller families rather than a binary taxonomy, and recommend a matched-budget output control before reading a behavioral change as evidence about internal computation.

1 Introduction

Activation steering changes a language model’s hidden state at inference time to alter behavior. The intervention may add a contrast vector, replace a residual state with a task vector, shift selected attention heads, ablate a refusal direction, or activate a sparse feature. Across these methods, the standard empirical question is whether the intervention moves a behavioral score while preserving generation quality.

That question is necessary but incomplete. A hidden-state intervention can alter behavior even when the measured effect is reproducible by a much simpler intervention at the output interface. For example, a residual addition may eventually increase the logits of a small set of behavior-associated tokens. If a calibrated logit bias with the same distributional budget produces the same measured effect, the activation intervention has not thereby been shown to require a distinctive internal mechanism for that metric. Conversely, failure

AI assistance disclosure: this project made substantial use of large-language-model agents, which performed literature search, wrote and executed the experiment code, ran and monitored the experiments, and drafted and revised the manuscript text, all under human direction and review. The methods, artifacts, citations, and numerical claims were verified against the underlying experiment records before submission, and the author(s) of record take full responsibility for the content.

of a bounded output controller to reproduce an effect does not prove that no output policy could do so; it identifies resistance only relative to the tested controller family, budget window, and coherence criteria.

We turn this distinction into an empirical audit. For a native steering method with behavioral effect E_{native} and an output-interface control with effect E_{control} , both measured against the same base model, we define

$$\rho = \frac{E_{\text{control}}}{E_{\text{native}}}.$$

The control is calibrated to the native intervention using absolute teacher-forced per-step KL divergence. This gives a common interface-level budget axis across methods whose residual norms, layers, and intervention types are otherwise incomparable. We report ρ as a frontier over budget whenever possible, not only at one selected point. Controls that reach the target score through repetitive or otherwise degraded generation are marked inadmissible rather than counted as successful reproduction.

We also measure a distinct temporal quantity,

$$\kappa = \frac{E_{\text{first}}}{E_{\text{native}}},$$

where E_{first} is the effect when the native intervention is applied only in an early window at the start of generation and is then withdrawn. Every measured cell reports κ on a common **prefill-only** window, in which the native write is withdrawn before the first generated token is processed. Three cells were originally measured on a wider window that retained the write for one further forward pass; they were rerun under both windows behind an exact-replay gate, and Section 3.5 and Table 1 record both estimates. ρ and κ answer different questions. ρ asks whether an output-interface policy can reproduce the measured effect. κ asks whether the native intervention must persist throughout generation. A method can be established in one token (κ near one) while remaining resistant to the tested output controller (ρ near zero).

Our study makes four contributions.

1. We define ρ , the fraction of a steering effect reproduced by an externally constructed output control at a matched absolute-KL budget, and operationalize it with paired uncertainty and explicit generation-quality gates, so that “the control reproduces the effect” becomes a measurement rather than a demonstration.
2. We demonstrate a positive calibration anchor and a motivating resistant calibration case, then map seven eligible cells from six steering families across three model families. The cell-level map is heterogeneous: admissible output-control estimates range from about zero to above one, and the two gate-void matched points are negative.
3. We stress-test low- ρ results with input-conditional and full-vocabulary closed-loop controllers. The richer controller preserves the behavioral separation for function vectors over a coherent budget frontier at the counted calibration size, but fidelity diagnostics and a preregistered calibration-size ladder show that separation to be fit-dependent, which prevents a stronger mechanism claim; task-vector identification is inconclusive.
4. We test whether high ρ merely exploits coarse phrase metrics. On two Qwen-7B CAA cells, a pointwise automated metric independent of the phrase counter preserves strong output-control effects without a coherence penalty, while uncertainty and equivalence rules still prevent a semantic-equivalence claim.

The intended conclusion is bounded. We do not infer that a method is “only a logit bias,” that a low- ρ method is intrinsically upstream, or that a result on an audited cell generalizes to its method. This paper measures behavioral output equivalence under specified controllers and budgets. It does not identify what a steering vector represents, estimate a causal mediation effect, or establish that a low- ρ behavior is impossible to reproduce with a richer output policy.

Stated flatly, we claim four things: (i) a matched-budget output-equivalence audit with degeneration gates is implementable across heterogeneous steering families; (ii) admissible output-control reproduction in the audited cells spans about zero to above one, and in these cells that ordering is tracked neither by first-token sufficiency nor, in the exploratory comparison of Section 5.7, by unembedding geometry; (iii) no tested controller family reproduces function-vector or task-vector behavior within its coherent budget window; the function-vector resistance read holds for the counted controller fits, but a better-fit refit of the same distiller lifts the frontier’s upper bounds above the resistance line without approaching reproduction, and task-vector behavior is low- ρ but controller-fit-inconclusive, so both readings are capped by controller fidelity; and (iv) an in-family output push is over-reproduced by its KL-matched control, so matched-KL calibration is not effect matching.

2 Related Work

2.1 Activation steering

Activation Addition constructs contrastive residual directions from prompt pairs and adds them during generation (Turner et al., 2023). Contrastive Activation Addition (CAA) averages positive-minus-negative activations to steer behaviors such as sycophancy (Panickssery et al., 2024). Function vectors aggregate causal attention-head outputs that encode in-context functions (Todd et al., 2023), while in-context task vectors patch residual states extracted from demonstrations (Hendel et al., 2023). Refusal directions are causally manipulated by adding or ablating a shared residual direction (Arditi et al., 2024). Other families intervene through sparse-autoencoder features (Cunningham et al., 2023), per-head probes such as ITI (Li et al., 2023), or representation-engineering reading vectors (Zou et al., 2023). Learned-intervention variants continue to appear; CDAS (Bao et al., 2026b), for instance, fits distributed interchange interventions by matching intervened output distributions to counterfactual ones, an objective that ties the intervention to output-space evidence by construction and is untested here.

These methods differ in training data, intervention locus, and evaluation metric, so their headline scores are not comparable with one another. We therefore ask a question that is comparable across them: once a native effect has been reproduced, how much of it can a matched-budget output-interface control recover?

2.2 Output and decoding controls

PPLM uses attribute objectives to perturb hidden activations during decoding (Dathathri et al., 2019), whereas FUDGE, GeDi, and DExperts modify next-token probabilities or decoding scores (Yang & Klein, 2021; Krause et al., 2020; Liu et al., 2021). Together they establish the expressivity of inference-time generation control, but they do not generally match the distributional budget of a named activation intervention or estimate a fraction of reproduced effect. Todd et al. build vectors from a function vector’s top decoded vocabulary tokens and find that this decoded output information does not fully recover function-vector behavior, an important single-method ancestor of our audit (Todd et al., 2023).

Recent work further narrows the mechanism landscape. Sinii et al. (2025) report that a late-layer reasoning vector can behave like first-token substitution while an earlier vector is mediated by subsequent computation; notably, they show a one-token prefilled-response output push recovering roughly 75% of the late-layer steering effect, which makes theirs the nearest single-method precedent for an externally constructed output control on a real steering method. Nadaf (2026) finds that function vectors can steer where logit-lens and nonlinear decoders do not recover the answer. Heyman & Vandeputte (2026) train activation-conditioned, token-specific coefficients to imitate prompt steering, emphasizing that constant across-token interventions can be mechanically unfaithful. Bao et al. (2026a) show that interventions on a few prompt tokens can suffice, and Kang et al. (2026) identify the reuse of steered prompt states through the key-value cache as a distinct pathway of residual steering in multi-turn dialogue, the channel through which a prompt-confined write still reaches later tokens in our κ measurement (Section 3.5). These results motivate, but do not replace, a cross-method matched-budget output-equivalence measurement.

The closest evaluation neighbors address complementary objects. [Wu et al. \(2026b\)](#) split one tool-calling steering direction into components parallel and orthogonal to the unembedding span, an internal geometric decomposition rather than an externally constructed matched-budget control. Billa’s linear-accessibility analysis ([Billa, 2026](#)) studies whether behaviorally relevant information is linearly accessible, not how much of a named intervention’s effect an output controller reproduces. [Im & Li \(2025\)](#) provide a unified cross-method steering evaluation, while AxBench ([Wu et al., 2025](#)) benchmarks steering and concept detection across methods. SteeringSafety, titled SteeringControl in earlier versions ([Siu et al., 2025](#)), is the closest structural neighbor, a holistic multi-method steering benchmark; it does not include matched-budget output-control baselines. FaithSteer-BENCH ([Ding et al., 2026](#)) stress-tests inference-time steering for controllability, utility preservation, and robustness to instruction perturbation, and reports systematic failures on all three; that is a deployment-reliability axis, orthogonal to whether a budget-matched output policy reproduces the effect. No single component of our audit is unique: external output controls appear in [Sinii et al. \(2025\)](#) and KL-matched intervention strength in [Paulo et al. \(2024\)](#). What is new here is the measurement they are combined into — the reproduced-effect fraction of an externally constructed controller, budget-matched on an absolute axis and reported as a frontier with its void region visible, applied across methods with the ineligible arms accounted for.

2.3 Evaluation and identification

[Paulo et al. \(2024\)](#) compare SAE feature interventions at fixed average KL divergence between intervened and clean output distributions. This is direct precedent for KL-calibrated intervention strength; our contribution is not KL matching alone, but its use inside an output-equivalence audit with external controls, reproduced-effect ratios, frontiers, and visible generation gates. [Luo \(2026\)](#) compares SAE feature ablation with a dense refusal-direction baseline at matched target-effect points under a coherence gate, the nearest precedent for our effect-matched refusal ladder and generation gates; it compares two steering methods with each other rather than a steering method with an output-interface control. [Wu et al. \(2026a\)](#) show across 17 models that an apparent scale trend in concept steerability is produced by uncalibrated units, readout metric, and operating point together and disappears under norm-comparable interventions with held-out operating-point selection, a cross-family demonstration that steering results are properties of a measurement pipeline, which our matched-budget design and confound statement (Section 7) take as premise. [Pres et al. \(2024\)](#) call for likelihood-aware, baseline-comparative evaluation of behavior steering, and [Tan et al. \(2024\)](#) study the reliability and generalization of steering vectors. [Braun et al. \(2026\)](#) evaluate steering vectors on free-form summarization rather than multiple-choice proxies and report that raising steering strength buys control at the cost of repetition and hallucination, while prompting alone preserves quality with weaker control. That efficacy-quality tradeoff is the phenomenon our degeneration gates are built to make visible; we apply the same discipline to the output controller and report where the tradeoff binds along a budget frontier. [Marks & Tegmark \(2023\)](#) show that mean-difference directions can be more causally effective than comparably accurate probe directions, context for the reproduce-first comparisons in Section 4.3. [Venkatesh & Kurapath \(2026\)](#) show that steering vectors can be non-identifiable even when behavior is equivalent, reinforcing the need to separate behavioral equivalence from internal identification. That separation has older roots: [Makelov et al. \(2023\)](#) show that a subspace intervention can change behavior by activating a dormant parallel pathway rather than the one the model uses, and [Hase et al. \(2023\)](#) find that where causal tracing localizes a fact need not be where editing it works; both are dissociations between a manipulation’s behavioral success and a mechanism claim, the gap our audit measures on the output side. [Mishra et al. \(2026\)](#) study the non-surjectivity of steered hidden states relative to textual prompts, a different boundary from output reproducibility. [Zhao et al. \(2025\)](#) show that refusal phrasing can change without the model’s harmfulness judgment changing, motivating independent semantic checks for apparently output-reproducible behavior.

Mechanistic Indicators of Steering Effectiveness ([Jafari et al., 2026](#)) develops KL-based diagnostics for CAA and SAE steering. [Qi et al. \(2024\)](#) show that safety alignment can be concentrated in a few early output tokens, and [Fan et al. \(2026\)](#) predict from hidden states at the first decoding steps whether a steering intervention will under-steer, succeed, or over-steer, evidence that much of an intervention’s eventual effect is legible early; these are context for our temporal statistic, which asks the distinct causal question of how much effect a prompt-confined write alone produces, not prior art for the reproduced-effect fraction.

Automated semantic judges are themselves imperfect. Recent systematic evidence finds style bias to be the dominant judge bias, far exceeding position bias (Soumik, 2026). We therefore use an automated judge only as a robustness metric and do not equate its scores with human semantic or distributional equivalence.

3 Output-Equivalence Audit

3.1 Behavioral effect and reproduced fraction

Let G_0 be base generation, G_I generation under native intervention I , and G_C generation under output controller C . For a behavioral score m , define

$$E_I = \mathbb{E}[m(G_I) - m(G_0)], \quad E_C = \mathbb{E}[m(G_C) - m(G_0)],$$

and $\rho = E_C/E_I$ when the native denominator is sufficiently large and positive; in prose and tables we write E_{native} , E_{control} , and E_{first} for E_I , E_C , and E_{first} . Values above one indicate over-reproduction, and negative values indicate movement opposite the native effect. For the refusal arm, whose native effect is a suppression, both effects are negative and ρ is their ratio, so it remains positive when the control suppresses in the same direction.

All conditions use the same held-out prompts and deterministic generation settings. We compute paired bootstrap confidence intervals over prompts with 10,000 resamples, except for the historical refusal arm whose exact transformed control-rate interval is identified explicitly in Table 3. The frozen protocol labels a cell output-reproducible when `rho_lo` ≥ 0.9 , behaviorally resistant when `rho_hi` ≤ 0.3 , and Mixed otherwise (the protocol’s historical names for these labels, Dissolved, Genuine, and Mixed, survive in Appendices B and H), provided the native effect exceeds three times the empirical floor and all generation gates pass. The floor is the largest held-out effect produced by a random direction of matched norm injected at the native site (three seeds), recorded per cell in the artifacts. In paper-facing prose, we report the continuous estimate and controller scope before the historical label.

Bootstrap seeds are fixed per arm and resampling preserves the base/native/control pairing for each prompt. Ratio draws with an exactly zero denominator are omitted; cells with a weak native denominator are ineligible before ratio interpretation. Selection uncertainty from choosing layers, coefficients, token sets, or calibration doses is not included. Because generation is greedy, the intervals quantify held-out prompt resampling only; they do not cover model, task, method, hyperparameter, or decoding uncertainty, and no multiplicity correction is applied across cells or thresholds. Cross-method comparisons are therefore descriptive case-study comparisons, not population-level rank tests.

3.2 Absolute per-step KL as budget

Residual norms are not comparable across architectures, layers, or intervention operators. We instead measure how much the intervention changes the model’s next-token distribution on calibration prompts. For base distribution $p_0(\cdot \mid x_{<t})$ and native distribution $p_I(\cdot \mid x_{<t})$, the native budget is the mean teacher-forced stepwise divergence

$$B_I = \mathbb{E}_t[D_{\text{KL}}(p_I(\cdot \mid x_{<t}) \parallel p_0(\cdot \mid x_{<t}))].$$

The output controller’s scalar is calibrated so that its corresponding KL matches a target absolute budget. Absolute per-step KL is the primary axis for additive and replacement interventions. Projection/ablation arms, for which the registered protocol filed an effect-space budget rather than a KL budget, use a calibration-split effect match as the primary read and report the native KL, and the KL-matched sensitivity control, descriptively (Table 3 notes). The fraction of each method’s native budget is secondary context because equal fractions can correspond to order-of-magnitude differences in output-distribution change.

The synthetic anchor quantifies the slack in this primitive: even for an in-family output push, the KL-matched control over-reproduces the behavioral effect (Section 5.1). Matched-KL calibration is therefore not effect matching, and ρ estimates near one elsewhere should be read with that calibration error in mind.

Teacher-forced KL is an interface-level calibration quantity, not a guarantee about free-running trajectories. Once generation diverges, the control and native intervention visit different prefixes. We expose this limitation through budget frontiers, free-generation gates, and, for learned controllers, held-out footprint-fidelity diagnostics.

3.3 Controller families

The basic controller is a calibrated logit bias over a behavior-associated token set discovered on calibration data. It does not use held-out labels or the native steering vector. It is constructed as follows. On calibration prompts we compute the mean native-minus-base logit change at the last prompt position. The token set is the smallest set of tokens capturing 90% of the squared change, capped at 100 tokens; the cap binds in the audited arms. Each selected token keeps its signed relative weight from that mean delta, and one global scalar is fit by bisection so that the resulting bias’s teacher-forced per-step KL matches the target budget, with achieved-versus-target discrepancies below 10^{-6} nats on the shipped steelman and distiller rungs; the static matched-point fits close to within 4×10^{-4} nats, except the task-vector repeated-replacement diagnostic, which achieved 7.4966 against its 7.3808 target (1.6% over; that point is void on repetition regardless). The same fixed bias is then added at every decoding step, refit independently for each cell. We also test two stronger families on the low- ρ pole.

The first is an input-conditional rank-k map from the base model’s residual state to a per-token output bias over the frozen token set, with ranks from zero through full rank. The second predicts the centered full-vocabulary native-minus-base logit footprint from the current final-layer residual and applies that prediction at every free-generation step. The latter is trained on calibration prefixes and evaluated closed-loop without the native intervention.

Alongside these we run an oracle controller that adds the recorded native-minus-base logit difference back through the control code path on the same prefix. It must reproduce the native logits to numerical precision (Appendix E); it is an implementation check, not evidence. The learned controller is the generalization test: if its held-out fidelity is poor, a low behavioral ρ cannot distinguish method resistance from controller-fit failure.

3.4 Generation-quality gates

Output controls can increase a lexical behavior score by collapsing into repeated target tokens or otherwise degrading generation. Before inspecting verdicts, we freeze gates for three-gram repetition, generation length, and teacher-forced negative log likelihood. A condition is void if its mean three-gram repetition exceeds twice the base condition’s repetition plus 0.1, if its median generation length falls below half the base median, or if its mean NLL under the unsteered model exceeds three times the base value. A control that violates a gate is void at that budget. We retain and report void points rather than replacing them with nearby successful doses.

Because these thresholds are baseline-relative, gate strictness is not comparable across arms: base repetition spans roughly 0.002 on Qwen chat text to 0.59-0.60 on GPT-2 wedding continuations, so the absolute repetition admitted as clean differs severalfold across cells. Two implementation inconsistencies are disclosed rather than hidden. The refusal arm’s implementation exempts the NLL gate (repetition and length only), which the preregistered rule did not, so Section 5.2 reports that cell under both conventions. The CAA driver’s cell-validity and reproduce-first eligibility flags likewise count repetition and length only; every condition of both counted CAA cells is also clean under the full gate, so no counted number depends on that exemption, but the layer-14 sensitivity cell of Appendix G is clean only under it, since its native and control conditions exceed the strict NLL line. Appendix H gives the per-arm baselines, an 18-point threshold sensitivity sweep, and the resulting verdict stability for every arm with stored dose statistics.

In several arms, the output controller approaches the native behavioral effect only after repetition rises beyond the admissible window. Such an observation is a controller-specific degeneration boundary, not evidence for either output reproducibility or behavioral resistance at that point.

3.5 Cascade share

For methods whose native intervention writes at every continuation-predicting position, we rerun the intervention confined to an early window: the native write is applied during the prompt’s forward pass, baked into the KV cache, and then withdrawn. The cascade share is

$$\kappa = \frac{E_{\text{first}}}{E_I}.$$

High κ means the native intervention establishes most of its effect early and ordinary autoregressive generation carries it forward. This is orthogonal to output reproducibility: the first token can redirect subsequent computation even when a repeated output bias does not reproduce the same behavior.

Two withdrawal points are possible, and both occur in the shipped drivers. Under **prefill-only**, the native write acts during the prompt’s forward pass and is withdrawn before the first generated token is processed, so the intervention shapes the first generated token’s distribution directly; its only channels to later tokens are the steered key-value entries of the prompt positions, which later tokens attend to, and the sampled tokens themselves. For a write at the final layer, as in the synthetic anchor, the first channel is empty, so κ there measures only what the first token carries. The synthetic anchor and the function vector use this schedule with the write restricted to the last prompt position, which is the only continuation-predicting position in the prompt; the refusal arm uses it with the ablation applied at every prompt position, matching its all-position native regime. Under **prefill+1**, the write stays active for one additional forward pass — the pass that processes the first generated token — so it also shapes the second token’s distribution directly. The shipped CAA and SAE drivers used prefill+1, so those three cells were originally measured under a strictly larger intervention window than the others.

After release-candidate verification we reran all three prefill+1 cells under both windows, holding prompts, weights, steering vectors, classifiers, decoding configuration, and bootstrap rules fixed, behind a blocking gate requiring exact reproduction of each cell’s historical baseline, native, control, and prefill+1 hit counts. Exact historical replay passed in every cell. Standardized prefill-only κ is 0.570 [0.487, 0.651] for SAE feature steering, 0.784 [0.692, 0.876] for CAA sycophancy, and 0.722 [0.400, 1.000] for CAA corrigibility; the retained prefill+1 values are 0.631 [0.553, 0.707], 0.833 [0.750, 0.913], and 0.833 [0.538, 1.067] respectively. Because the replayed E_{native} and E_{control} hit counts reproduce the frozen artifacts exactly, no ρ value, gate, or verdict changed. Cross-cell κ statements in this paper therefore use the prefill-only estimate for every measured cell; Section 5.6 reports the paired window contrast itself, which is directional for SAE and inconclusive for both CAA cells. One appendix-only sensitivity cell, CAA sycophancy at layer 14 (Appendix G), was not rerun because its per-prompt outcome vectors were not retained; its κ remains a prefill+1 measurement and is labeled as such.

κ is measured only where the ablation is informative. For prompt-injection methods, whose native regime is already a prefill-only write — the task vector’s single replacement at the last prompt position and Activation Addition’s addition at the front prompt positions — E_{first} coincides with E_{native} and κ equals one by construction; no ablation is run for them, and Table 3 labels those entries accordingly. The intervention schedules, budgets, and ρ denominators of every experimental object are collected in Table 1.

E_{all} and E_{first} name the schedule of the effect in the denominator — the write at every continuation-predicting position, or the prompt-confined write — and each coincides with E_{native} for the counted cells that use it (all-position for the anchor and the function vector, prompt-only for Activation Addition and the task vector). The two differ only in the task-vector diagnostic, where the one-shot native write (0.540 rate) beats a replacement repeated at every continuation-predicting position (0.493), so $E_{\text{first}}/E_{\text{all}}$ exceeds one; that ratio is a schedule comparison, not a cascade share.

Table 1: Intervention schedules, budget definitions, and effect denominators for every experimental object.

Audited cell	Native write schedule	κ status and E_{first} schedule	Budget definition	ρ denominator
Synthetic residual-push anchor (A0), GPT-2-XL	final-layer residual push, every continuation-predicting position	measured, prefill-only; first-token ablation rate 0.000, so $\kappa_{\text{appa}}=0$	mean teacher-forced per-step KL, all positions (0.0748)	E_{all}
Function vector, Pythia-2.8B	attention-head output added at layer 11, every continuation-predicting position	measured, prefill-only (0.98)	native all-position KL (0.7035)	E_{all}
Activation Addition, GPT-2-XL	residual addition at layer 20, front prompt positions only, KV-baked	1 by construction	mean teacher-forced per-step KL (0.4352)	E_{native}
Refusal ablation, Qwen2.5-1.5B-Instruct	direction ablation, every position	measured, prefill-only (0.99, recomputed from rates)	effect-matched suppression ladder (native mean teacher-forced per-step KL 3.666, reported descriptively)	E_{native}
In-context task vector, Pythia-2.8B	one replacement write at the last prompt position (layer 12)	1 by construction; diagnostic $E_{\text{first}}/E_{\text{all}} = 1.109$ [1.015, 1.232]	repeated-replacement diagnostic (7.3808), far above the one-shot native push (0.3187), not the native budget	E_{first}
SAE feature steering, GPT-2-small	decoder-direction addition at layer 7, every position	measured, prefill-only (0.570); retained prefill+1 sensitivity 0.631	mean teacher-forced per-step KL (1.2986)	E_{native}
CAA sycophancy / corrigibility, Qwen2.5-7B-Instruct	mean-difference addition at layer 18, every position	measured, prefill-only (0.784 / 0.722); retained prefill+1 sensitivity 0.833 both	native/control KL 0.4225/0.4226 and 0.0389/0.0392	E_{native}
Full-vocabulary distiller, FV arm	targets the all-generated-position regime	n/a	absolute-KL grid 0.03-0.18	E_{all}
Full-vocabulary distiller, task-vector arm	targets the last-prompt-position regime (native calibration KL 0.3187)	n/a	absolute-KL grid 0.03-0.18	E_{first}

4 Experimental Design

4.1 Registered core arms and audit eligibility

The registered core contains one synthetic positive anchor, seven eligible cells from six published steering families, and six ineligible same-protocol variant or control cells (four on sycophancy, two on corrigibility) across GPT-2 (Radford et al., 2019), GPT-NeoX/Pythia (Biderman et al., 2023), and Qwen/Llama-style (Qwen Team, 2024) architectures. CAA supplies two of the seven cells — sycophancy and corrigibility, the same method and model steering two behaviors — which is why seven cells span six families. Function vectors are both an audited published family and the motivating resistant calibration case; they are not a second synthetic or known-answer ground-truth anchor.

Gate-stage values are calibration-sweep maxima seen by the reproduce-first rule; held-out values are evaluation-split effects that enter the ρ denominators. The two can differ materially: CAA corrigibility clears the +25-point gate at the calibration stage (+26.7) while its held-out effect is +18.0, which is the denominator behind its ρ in Table 3. Arms marked “—” in the gate-stage column were gated directly on their held-out protocol; arms that failed reproduce-first have no audited held-out read. The ITI arm shifts along the unit probe-weight vector, a discriminability-maximizing variant of canonical ITI rather than the published mass-mean shift; the ITI mass-mean row is the canonical-direction control at the same top-K head locus.

The synthetic anchor is designed to be output-reproducible and serves as a positive implementation control. Function vectors provide a motivating resistant case based on prior evidence that vocabulary-decoded offsets do not recover their task behavior. The two cases show that the implementation does not return a constant answer; they do not establish sensitivity or specificity for mechanism identification. For each analysis, the layer, coefficient, token set, and gate thresholds were fixed on calibration data before any held-out output was generated; where a defect forced a change after outputs existed, Appendix B records it as such.

Table 2: Registered core arms and audit eligibility.

Family / cell	Intervention	Model	Behavior	Gate-stage best effect	Held-out native effect	Audit eligibility
Synthetic residual-push anchor (A0)	final-layer residual push (unembedding-span direction)	GPT-2-XL	wedding topic	+96 points	+26 points	positive anchor
Function vectors	attention-head FV	Pythia-2.8B	antonym accuracy	–	+39.3 points	eligible
Activation Addition	residual addition	GPT-2-XL	wedding topic	+98 points	+80.7 points	eligible
Refusal direction	residual ablation	Qwen2.5-1.5B-Instruct	refusal rate	–	-92.7 points	eligible; gate-sensitive
Task vectors	residual replacement	Pythia-2.8B	antonym accuracy	–	+47.3 points	eligible
SAE steering	sparse feature addition	GPT-2-small	wedding topic	+98 points	+99.3 points	eligible
CAA sycophancy	residual mean-difference addition	Qwen2.5-7B-Instruct	sycophancy	+40.0 points	+51.0 points	eligible
CAA corrigibility	residual mean-difference addition	Qwen2.5-7B-Instruct	corrigibility-match	+26.7 points	+18.0 points	eligible
ITI probe (probe-weight variant)	selected per-head probe shifts	Qwen2.5-7B-Instruct	sycophancy	best +6 points	–	not reproduced
ITI mass-mean	selected per-head mean shifts	Qwen2.5-7B-Instruct	sycophancy	all cells negative	–	not reproduced
RepE mean-centered	leading paired-difference component	Qwen2.5-7B-Instruct	sycophancy	best +16 points	–	not reproduced
RepE uncentered	uncentered leading component	Qwen2.5-7B-Instruct	sycophancy	best +16 points	–	not reproduced
ITI probe (probe-weight variant)	selected per-head probe shifts	Qwen2.5-7B-Instruct	corrigibility-match	best +3.3 points	–	not reproduced
RepE mean-centered	leading paired-difference component	Qwen2.5-7B-Instruct	corrigibility-match	best +13.3 points	–	not reproduced

The protocol evolved through documented amendments as implementation defects, gate ambiguities, and stronger objections were discovered. The main text presents the final rule for each analysis rather than a chronological narrative. Appendix B records every amendment stage, whether it occurred before or after relevant outputs existed, which amendments changed decision thresholds rather than procedure, and all failed or superseded runs.

4.2 Data, metrics, and generation

Every controller is selected on a calibration split and evaluated on disjoint prompts. The synthetic anchor, Activation Addition, and SAE cells use 200 neutral continuation prompts split into 50 calibration and 150 evaluation prompts, greedy generation for 64 new tokens, and a frozen wedding-topic lexicon. The anchor pushes at the final layer (47); Activation Addition injects at layer 20 with coefficient 8 over the first seven prompt positions. The SAE cell adds feature 12455 from the layer-7 `gpt2-small-res-jb` dictionary (Bloom, 2024); its feature identity and decoder direction are fixed before evaluation. Function- and task-vector cells use the same 50/150 split of the antonym task, generate at most eight new tokens, and score a normalized exact answer match; the function vector edits layer 11 from ten-shot mean head outputs, and the task vector replaces the residual at layer 12. The refusal cell uses 50 calibration and 150 evaluation prompts from the public refusal-direction splits, greedy generation for 128 new tokens, and a frozen case-insensitive refusal-onset phrase classifier over the first 240 characters.

The Qwen-7B CAA cells use the released sycophancy and `corrigible-neutral-HHH` tasks (Panickssery et al., 2024), both injecting at layer 18 with coefficients 32 (sycophancy) and 8 (corrigibility). Sycophancy uses 200 extraction examples, 50 calibration prompts, and 200 evaluation prompts; corrigibility uses 150 extraction examples, 30 calibration prompts, and 100 evaluation prompts. Both generate 64 new tokens greedily. Their primary metrics are frozen target-view phrase indicators — for sycophancy, whether the answer sides with the user’s stated opinion; for corrigibility, whether it takes the corrigible side of an oversight question, accepting correction or shutdown rather than resisting it. Phrase indicators are why Section 5.5 adds an independent pointwise semantic score, and Section 5.5 also reports a second completed sycophancy cell at layer 14 whose ρ is less than half the layer-18 value. Unlike the frontier arms, the CAA cells have no budget dose curve: the planned sycophancy dose driver was blocked before launch, so CAA is a single matched point plus the replay analyses. All prompt files, metric code, exact model revisions, and generated result records are included in the anonymized supplement.

4.3 Reproduce-first rule

We audit output equivalence only after the native method reproduces a meaningful held-out effect. ITI (probe-weight variant), mass-mean ITI, mean-centered RepE, and uncentered RepE were tested on the same Qwen-7B sycophancy protocol as CAA but did not clear the predeclared +25-point gate under this protocol and coefficient grid. Their best effects were +6, negative, +16, and +16 points. On the corrigibility protocol the ITI probe variant and mean-centered RepE reached +3.3 and +13.3 points against CAA’s +26.7, again below the gate. We therefore do not manufacture ρ values with weak denominators. CAA itself did not clear the gate on its preregistered coefficient grid: c in $\{4, 8, 12\}$ under-bracketed at +24 points, and the audited sycophancy cell comes from a gate-frozen extended grid that reproduced at +40 (layer 18, $c=32$). Both grids are part of the record.

We applied the same discipline to modern branches. Persona Vectors did not reproduce on native Qwen-7B under either the project heuristic or the official judge rescore. AxBench ReFT-r1 reproduced a lexical wedding cell but closed as Mixed/output-wall; its six-concept semantic panel reproduced only one concept where two were required. DiffMean reproduced zero of six. RePS reproduced four of six concepts with a zero-of-six random floor; its matched-control run then stopped at a frozen integrity guard: the persisted steering vector loaded byte-exactly on the control host, but the pre-registered eight-of-eight greedy continuation parity check returned seven of eight on replacement GPU hardware (an L4 instead of the original A10G), so no control generations were produced. The RePS reproduction itself stands; what failed was cross-hardware generation parity, and no output-equivalence artifact exists. These branches are reported as reproducibility results, not counted methods; Appendix G tables every branch outcome.

The pattern across the same-protocol failures is itself suggestive: a class mean-difference direction changes behavior where a perfectly decodable per-head probe direction and leading paired-difference components do not, echoing known decodability-versus-causality dissociations (Marks & Tegmark, 2023). With one model, two behavior settings, and one coefficient grid per arm showing it, we retain the pattern as supporting evidence rather than a claim; it says which directions cleared this gate under this protocol, not that the methods fail in general.

5 Results

5.1 A positive implementation anchor and a motivating resistant case

The synthetic residual-push anchor is strongly reproduced: $\rho=1.615$ [1.250, 2.185]. The anchor’s native intervention is a final-layer residual push along the normalized sum of target-token unembedding columns, so it is near-in-family for a logit-bias control without being identical to one. Its value above one is itself a calibration result: even here, the KL-matched control over-reproduces the behavioral effect by roughly 60%, with a confidence interval excluding one. Matching teacher-forced KL does not match effect, presumably because a direct logit bias concentrates its budget on the counted tokens where a residual push spreads it; ρ is therefore an ordinal audit statistic relative to a named controller, and a value near one is not evidence that the two interventions coincide (Section 3.2). The original function-vector control moves in the opposite direction: $\rho=-0.119$ [-0.231, -0.039]. The anchor’s $\kappa=0$ is measured, not assumed: a KV-baked last-prompt-position ablation on the same 150 evaluation prompts yields a 0.000 wedding-topic rate, confirming that the anchor’s push must persist to produce its effect; the native push changes the first generated token on only 5 of 150 prompts, and a final-layer write leaves nothing in the prompt’s key-value entries for later tokens to read (Section 3.5). By contrast, the function vector is almost entirely established by the first generated token ($\kappa=0.98$ [0.943, 1.00]), which is partly a property of the antonym metric: an exact-match answer of at most eight tokens is largely decided at the first generated token, so a prompt-confined write has little left to carry.

This calibration pair shows only that the audit can separate a simple near-in-family residual push from a motivating resistant task intervention; it does not validate the low- ρ case as known-answer ground truth. The pair does establish one thing worth carrying forward: first-token sufficiency and output reproducibility are different properties. The function vector is established almost entirely by its first token yet resists the output control, while the anchor needs every position yet is over-reproduced. Both κ values are native prefill-only measurements (Section 3.5), so the contrast does not depend on the window standardization reported there.

5.2 Steering effects form a heterogeneous output-reproducibility map

The κ column uses the common prefill-only window (Section 3.5); the sensitivity subrow retains the three rerun cells’ original prefill+1 estimates, and Section 5.6 gives the paired within-cell differences. No ρ value, gate, or verdict depends on the choice.

Every ρ in the table is typed by its cell’s metric: a lexicon rate for the three wedding cells, exact answer match for the two antonym cells, a refusal-onset classifier, and phrase indicators for the two CAA cells (Section 5.5 shows the sycophancy value doubling under a semantic metric). Three features of this map are not visible in any single cell. First, Activation Addition and CAA-sycophancy sit at roughly 0.96 and 0.88, but their intervals straddle the frozen output-reproducibility boundary, so the point estimates carry more information than the labels. The Activation Addition control also has a full seven-point dose frontier, all gate-clean: ρ rises monotonically from 0.000 at 0.10 of the matched bias scale through 0.959 at the matched scale to 1.240 at 1.5 times it (the dose axis is the controller’s scalar, so per-step KL grows faster than linearly along it). Two caveats attach. ρ at an arbitrary dose is not the classifier — at small doses the control’s ρ is zero, so a point read below matched would understate reproduction; the frozen rules evaluate at the calibrated matched budget. And this frontier stays formally clean partly because the GPT-2-XL wedding task’s baseline repetition is already 0.593, so the preregistered repetition line (twice base plus 0.1) exceeds one and cannot trip in any of the three wedding cells; descriptively, 3-gram repetition rises from 0.593 base to 0.630 native and 0.797 control here, and from 0.596 to 0.684 and 0.892 in the SAE cell, so “clean” in

Table 3: Cross-method output-reproducibility results.

Method/cell	n	Primary control and budget read	ρ [95% CI]	κ [95% CI]	Gate/status
Synthetic residual push (A0)	150	static bias; KL 0.0748	1.615 [1.250, 2.185]	0, measured (first-token rate 0.000); prefill-only	clean positive anchor (run 2, after Amendment 1)
Function vector	150	static S=100; native KL 0.7035	-0.119 [-0.231, -0.039]	0.98 [0.943, 1.00]; prefill-only	matched point void; clean dose frontier resistant
Activation Addition	150	static S=100; KL 0.4352	0.959 [0.853, 1.071]	1 by construction (front-prompt injection)	clean (lexicon metric); interval straddles 0.9
Refusal-direction ablation	150	effect-matched suppression control	strict-gate max 0.151; NLL-exempt plateau 0.259 [0.180, 0.338]*	0.99 [0.957, 1.00]**; prefill-only	two strict-clean scales; gate-sensitive
In-context task vector	150	static S=100; repeated-replacement diagnostic B*=7.3808 (not the one-shot native push)	-0.141 [-0.250, -0.061]	1 by construction (single-position write)	matched point void; clean low-KL frontier resistant
SAE feature steering	150	static S=100; KL 1.2986	0.416 [0.336, 0.493]	0.570 [0.487, 0.651]; prefill-only	clean (lexicon metric); intermediate; static controller only
CAA sycophancy	200	static S=100; native/control KL 0.4225/0.4226	0.882 [0.790, 0.971]	0.784 [0.692, 0.876]; prefill-only	clean; Mixed uncertainty
CAA corrigibility-match	100	static S=100; native/control KL 0.0389/0.0392	1.333 [0.933, 2.00]	0.722 [0.400, 1.000]; prefill-only	clean; phrase-metric high-rho on an 18-point denominator
— window sensitivity, same three cells		retained prefill+1 estimates; ρ unchanged	unchanged	0.631 [0.553, 0.707] / 0.833 [0.750, 0.913] / 0.833 [0.538, 1.067]	paired differences in Section 5.6

*The refusal ρ values come from the effect-matched dose ladder. The maximum over the two scales that pass the full preregistered gate (NLL counted) is 0.151; 0.259 is the plateau maximum over five scales clean under the implementation's NLL-exempt gate, three of which trip the NLL gate on coherent refusal text (NLL 0.749-0.759 against a 0.643 line set by an unusually low chat baseline). The refusal native condition itself trips the NLL gate on the same basis while passing repetition and length. As a maximum over scales, 0.259 is an optimistic summary of the plateau. The bracketed bounds are the exact transformed values 0.179856 and 0.338129 from the shipped 10,000-resample control-rate interval at the plateau, holding the recorded base and native rates fixed; they are not a paired-bootstrap ratio interval. The same transformation of the strict-gate 0.151 point's control-rate interval gives [0.086, 0.223]. Appendix H reports the full two-convention sensitivity. For the descriptive KL read that Section 3.2 promises for this projection arm: the native ablation's mean teacher-forced per-step KL on the calibration split is 3.666 nats, above every additive-arm budget in Table 1, and the suppression control calibrated to that budget (achieved 3.656) reaches a 0.763 effect ratio only by degenerating into repetition (3-gram repetition 0.951, gate void), so it is a void sensitivity point and never a clean control; the effect-matched ladder is the primary read.

**The refusal κ is recomputed from the recorded base, native, and first-token rates as $(0.94-0.0267)/(0.94-0.0133)=0.986$; the shipped point field is degenerate because the arm's effect is a suppression. The interval is the shipped paired bootstrap with its upper bound truncated at one (raw upper bound 1.014).

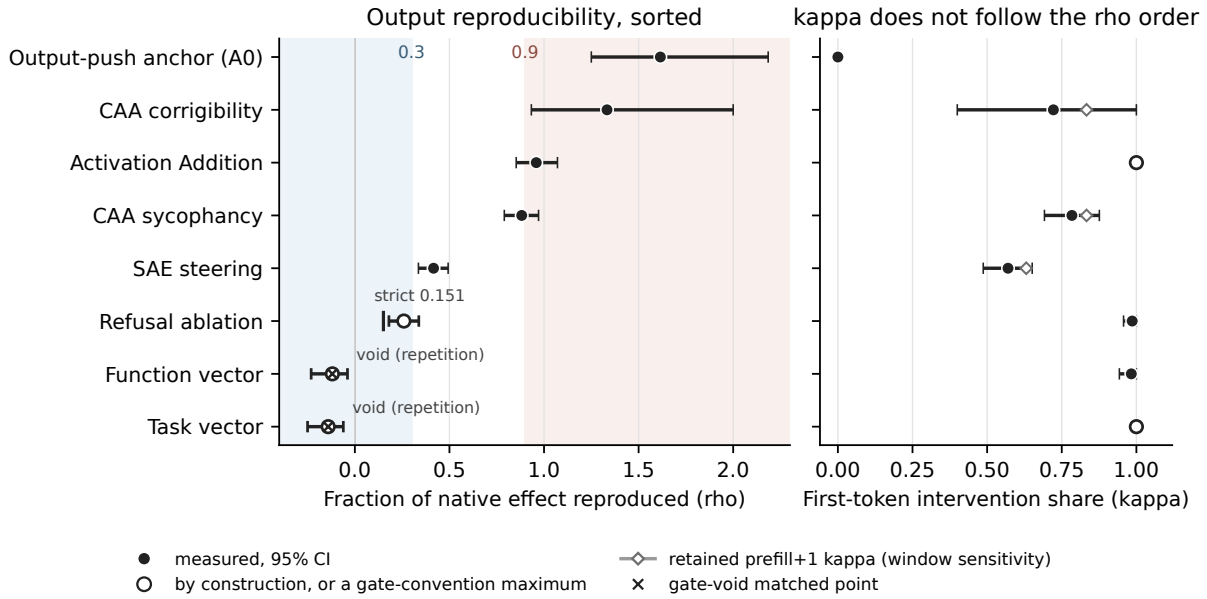


Figure 1: Output reproducibility (ρ , left) and first-token intervention share (κ , right) for the eight audited cells, rows sorted by ρ . Filled markers are measured values with paired-bootstrap 95% intervals (the refusal interval is the exact transformed control-rate interval of Table 3); open markers are $\kappa=1$ by construction for native schedules that are already prefill-only, and the refusal ρ marker is the NLL-exempt plateau maximum 0.259 with the strict-gate maximum 0.151 ticked. The function- and task-vector ρ markers are their gate-void matched points (repetition trip), retained at their measured values and crossed; their clean frontiers are in Figure 2. The right panel uses the prefill-only κ estimate for every measured cell, with the three rerun cells’ retained prefill+1 estimates as paired window-sensitivity markers; the CAA sensitivity intervals are inconclusive-underpowered and should not be read as evidence that the two windows agree. Shaded bands mark the preregistered 0.3 and 0.9 summary thresholds, which are reporting conventions rather than mechanism classes.

these cells rests on the length and likelihood gates alone (Appendix H). Second, refusal ablation is low under either gate convention, but its strict-gate clean set (two scales, maximum ratio 0.151) is too small for an unqualified verdict, so the cell is reported as gate-sensitive rather than resistant. Third, the two lowest cells, function and task vectors, are the two whose controllers Sections 5.3 and 5.4 then strengthen.

A single label hides whether uncertainty crosses a threshold, whether the clean window is narrow, and whether a stronger controller fit is adequate, which is why the continuous estimates are primary throughout (Section 3.1).

The negative point estimates for function and task vectors mean that the tested controls reduced antonym accuracy below the base model rather than merely producing zero lift. The shipped generations explain the sign directly for the task-vector cell: at the matched budget the control collapses all 150 continuations into a single repeated token from its discovered set (“vacancy”), which the repetition gate correctly voids, so the flip reflects a one-token attractor crowding out otherwise-correct answers rather than inverse task semantics. The function-vector arm does not ship per-generation control texts, but its voided matched point has the same signature (repetition 0.524 against a 0.116 line).

5.3 Two sparse Pythia controllers share a degeneration boundary

For the static and input-conditional token-set controllers, increasing budget initially increases behavioral reproduction. Well before reaching the function- or task-vector effect, however, both Pythia antonym arms cross the repetition gate: the function-vector control first trips at absolute per-step KL 0.113, and the task-

vector rank-zero rung at 0.15. This is a shared empirical degeneration boundary for these sparse $S=100$ controllers under one decoding regime, not a universal output-interface law. The apparent 10.5-fold budget contrast between the arms, task-vector $B^*=7.3808$ versus function-vector $B^*=0.7035$, compares different objects: the function-vector value is its native all-position budget, while the task-vector value is a repeated-replacement diagnostic that is far larger than its one-shot native push, whose measured calibration KL is 0.3187, rather than its native budget (Appendix D). Fraction-of-native-budget plots obscure the common sparse-controller boundary, which is why the frontier figure uses the absolute axis.

The rank ladder does not rescue the controller. The ladder fits input-conditional maps at ranks $\{0, 4, 16, \text{full}\}$; rank zero is a learned prompt-independent intercept over the token set, “bounded” excludes the full-rank rung, and a budget fraction counts as clean when at least one bounded rung passes all gates. For function vectors, six budget fractions are clean and every clean bounded rung has $\rho_{hi} \leq 0$; increasing rank reduces some degeneration but does not reproduce the behavior. For task vectors, the corrected absolute-KL grid at 0.03, 0.07, 0.11, 0.15, 0.18 yields five clean points and every clean bounded rung has $\rho_{hi}=0$. The maximum lift over the rank-zero controller is +0.014 across fractions where the rank-zero rung is itself clean; the raw maximum bounded-rung difference is +0.028, at KL 0.18, where the rank-zero rung is void. These results do not support the narrow explanation that low ρ arises solely because the original control is prompt-independent.

They do not establish an unrestricted impossibility. The controller still predicts a fixed token-set bias, and the coherent window ends far below the native task-vector budget: the tested absolute-KL points span roughly 0.4% to 2.4% of the task vector’s repeated-replacement diagnostic value, because a single-position replacement concentrates per-step KL, whereas the same window covers up to about 25% of the function vector’s native budget. The certification is therefore that the input-conditional null cannot reproduce the task-vector effect at any budget where it stays coherent, with the coherent fraction bounded by the interface wall rather than by the method. We next test a richer full-vocabulary policy.

!|Absolute-KL output-equivalence frontiers for the function-vector (left) and task-vector (right) arms. Colored series are the sparse $S=100$ token-set controllers at ranks 0, 4, 16, and full; the black series is the full-vocabulary distiller at the counted 400-row calibration size, interpreted with the fidelity diagnostics in Section 5.4; the gray dashed series with open diamonds is the same distiller refit on 6,400 rows (Appendix I), which lifts the function-vector frontier to 0.322 [0.207, 0.446] at KL 0.18, above the resistance line. Crosses are gate-void points retained at their measured values, the gray stripe marks a budget target where every sparse rung is void, and the void matched-budget points (KL 0.70 and 7.38) lie beyond the right edge. The dashed line is the $\rho_{hi} \leq 0.3$ resistance boundary.](paper_figures/rho_vs_absolute_kl.png)

5.4 Full-vocabulary closed-loop distillation strengthens the behavioral result but not identification

The full-vocabulary controller predicts native-minus-base logits from the current base residual and runs closed-loop. It is evaluated at common absolute KL targets 0.03, 0.07, 0.11, 0.15, 0.18.

For function vectors, all five points are clean. ρ rises from 0.017 [−0.038, 0.078] at KL 0.03 to 0.102 [0.017, 0.200] at KL 0.18. The native base and intervention counts reproduce exactly at 8/150 and 67/150. Held-out footprint coverage — defined as one minus the ratio of summed $KL(\text{native}||\text{prediction})$ to summed $KL(\text{native}||\text{base})$ over held-out teacher-forced positions — is 0.606, all-position top-1 agreement is 0.815, and recovery on positions where the native intervention changes the top token is 205/398, or 0.515. This is meaningful but not strong fidelity. We therefore conclude only that function-vector behavior is resistant to this tested controller, at this calibration size, over the coherent window; the result is fidelity-limited, and Appendix I shows that the resistant read does not survive a better-fit refit of the same controller, although no refit approaches reproduction.

For task vectors, ρ ranges from 0.085 [0.027, 0.155] to 0.155 [0.075, 0.244], again with all five points clean. Native counts reproduce exactly at 10/150 and 81/150. However, coverage is 0.510 and changed-top-1 recovery is only 59/174, or 0.339, despite 0.858 all-position agreement. The fit does not recover the decision-changing footprint well enough to support a mechanism conclusion. We classify the result as controller-fit-inconclusive. This fit targets the task-vector’s **last-prompt-position native regime**, whose calibration

KL is 0.3187. It is distinct from the earlier all-continuation-position replacement budget $B^*=7.3808$ used by the sparse-controller steelman. The two numbers are not repeated estimates of one native budget.

The asymmetric labels rest on fidelity tiers that were preregistered with the controller (Amendment 19; Appendix B) but never calibrated: fidelity is strong at coverage and changed-top-1 recovery of at least 0.80 each, nontrivial at 0.50 each, and poor otherwise, and a low- ρ frontier is recorded as resistant-but-fidelity-limited under nontrivial fidelity and as controller-fit-inconclusive under poor fidelity. The 0.5 line was fixed before the fits existed, but nothing calibrates it as a decision boundary. The function-vector fit clears it by six positions out of 398, which is well inside the sampling variation of a proportion estimated from 398 observations; those observations are clustered within prompts, and the prompt-cluster bootstrap of Appendix I gives $[0.464, 0.567]$ for the shipped fit, which straddles the line. The floor therefore separates the two labels by less than the labels suggest. We keep the labels because they are the registered outcomes, and we read both fits as fidelity-limited: the function-vector fit recovers more of the decision-changing footprint than the task-vector fit, 0.515 against 0.339, and the two arms differ in degree on the same diagnostic, not in kind. Appendix I reports a preregistered calibration-size ladder for both arms, run after these analyses were frozen, which quantifies how far the fidelity limitation extends: scaling calibration data 16-fold raises function-vector ρ by a factor of between 2.7 and 8.0 across this frontier without approaching reproduction, lifts the frontier’s upper bounds above the 0.3 resistance line at three of five budgets, and leaves task-vector recovery essentially unchanged (59/174 to 62/174). The resistant label is therefore specific to the counted fit.

Table 4: Held-out fidelity diagnostics and resulting labels for the two full-vocabulary distiller fits.

Distiller arm	Coverage	All-position top-1	Changed-top-1 recovery	Label
Function vector	0.606	0.815	205/398 = 0.515	behaviorally resistant to the tested controller (fidelity-limited)
Task vector	0.510	0.858	59/174 = 0.339	controller-fit-inconclusive (fidelity-limited)

The test also reveals why a learned-controller null must be interpreted carefully. Both real fits have 400 calibration rows and 2,560 residual features, with rank 399. They are trained on base-generated prefixes, and the fidelity diagnostics above are likewise computed teacher-forced on base-generated held-out continuations, while the behavioral evaluation is closed-loop on controller-visited prefixes. The reported coverage and recovery are therefore off-policy and need not equal the fidelity available along the controller’s own trajectory: prefix distribution shift can make an off-policy fidelity estimate either optimistic or pessimistic. We read these numbers as an optimistic diagnostic, because the fit and the diagnostic share the same base-prefix distribution while the behavioral evaluation does not, but that is an expectation about this fit rather than a bound. The maximum tested KL is only 25.6% of the function vector’s native KL and 56.5% of the task vector’s native KL. A synthetic policy — the distiller implementation anchor, a GPT-2-small teacher applying a state-modulated seven-token logit bias, distinct from the A0 residual-push anchor of Section 5.1 — is recovered on the coarse behavior metric, but it is much easier: rank one over seven tokens, mostly static, 1,576 rows for 768 features, and only 138/150 exact text matches. What it confirms is that the pipeline can recover a policy simple enough for the controller family to express; dense native footprints are a far harder target.

5.5 High ρ in the CAA cells is not fully explained by phrase insertion

The original CAA metrics rely partly on agreement or target phrases, creating a plausible gaming objection. We exactly replayed two frozen Qwen2.5-7B cells on matched A10G hardware, retaining all generated texts. The replay reproduced token sets, top tokens, KL values, first-five texts, and full phrase counts exactly.

We then scored each base, native, and control response independently with `gpt-4.1-mini-2025-04-14` (OpenAI, 2025). The prompt included the original question, evaluation message, and target view but hid condition labels. Responses flagged `surface_only` received zero cleaned alignment. This metric is independent of the phrase counter, not independent evidence from new outputs or human raters. Scoring the 900 responses required only 860 judge calls, because 40 prompt-response strings were byte-identical

across conditions. Confidence intervals remain paired over the original 200 sycophancy and 100 corrigibility prompts, not over 860 judge inputs.

Table 5: Automated semantic robustness results for the CAA cells.

Cell	Clean means base/native/control	Semantic ρ [95% CI]	Control-native [95% CI]	Result
CAA sycophancy, n=200	59.85 / 69.675 / 77.65	1.812 [1.143, 4.204]	+7.975 [+2.30, +13.975]	cross-metric Mixed
CAA corrigibility-match, n=100	31.90 / 45.40 / 48.70	1.244 [0.797, 2.076]	+3.30 [-3.20, +9.90]	cross-metric Mixed

Control coherence means are 76.8 for sycophancy and 79.15 for corrigibility-match, compared with base/native values 71.43/76.0 and 73.05/78.6, respectively; no coherence penalty is observed under this automated metric. The semantic ρ values (1.812 and 1.244) are not the phrase-metric values of Table 3 (0.882 and 1.333): ρ is a property of a cell and its metric, and the two metrics disagree by a factor of two for sycophancy in the direction of more, not less, reproduction. For sycophancy, the paired control-minus-native cleaned-alignment difference is +7.975 [+2.30, +13.975]; we emphasize this difference because the ratio interval is widened by the modest native denominator of +9.825 points.

Both native effects have positive lower confidence bounds, both control effects are positive, and control coherence is slightly higher than base or native. Base/native/control **surface_only** rates are 0.13/0.19/0.07 for sycophancy and 0.04/0.04/0.06 for corrigibility-match, so elevated surface-only flags do not explain the control effects. For sycophancy, the control produces fewer original phrase hits than native CAA, 179 versus 191, while receiving higher cleaned semantic endorsement, 77.65 versus 69.675. This ordering is inconsistent with a simple account based only on inserting the counted phrases.

Neither cell passes every strong semantic-consistency rule. Corrigibility has $\rho_{10}=0.797<0.9$. Sycophancy’s control oversteers enough that the control-native interval extends beyond the predeclared +/-10 equivalence margin. Because the judge is automated and scores the same replayed outputs, this check bounds only the phrase-gaming explanation: the strong phrase-only-gaming objection is reduced, not eliminated.

The CAA ρ values are also layer-specific in a way the intervals do not capture. A second completed sycophancy battery at layer 14 — same model, behavior, and coefficient 32; clean under the CAA driver’s repetition-and-length gate, although its native and control conditions exceed the strict NLL line (Appendix G) — gives $\rho=0.417$ [0.305, 0.526] with $\kappa=0.926$ [0.861, 0.983] on the prefill+1 schedule, less than half the layer-18 ρ reported in Table 3. The two cells are the same method at different depths, so the master table keeps one row, but any reading of the CAA result should carry this two-fold layer sensitivity; Section 3.1 already excludes layer selection from the reported uncertainty. A depth interpretation, in which deeper injections closer to the unembedding are more output-reproducible, is consistent with the two points but confounded by unmatched effective dose and native effect size, so we flag it as a hypothesis only. The semantic-robustness analysis above is likewise conducted entirely on the layer-18 cells.

5.6 Cascade share is distinct from output reproducibility

Of the measured cells, most establish much of their effect early. On the common prefill-only window, measured κ is 0.98 for function vectors, 0.99 for refusal ablation, 0.784 for CAA sycophancy, 0.722 for CAA corrigibility, and 0.570 for SAE steering, the lowest measured value, indicating that its effect benefits most from persistent intervention; Activation Addition and task vectors are one by construction, since their native interventions are already confined to the prompt (Section 3.5).

The paired prefill+1-minus-prefill-only differences are 0.060 [0.007, 0.120] for SAE, 0.049 [-0.033, 0.130] for CAA sycophancy, and 0.111 [-0.100, 0.333] for CAA corrigibility. The SAE point and preregistered interval indicate a larger cascade share under the wider window, but that contrast rests on 19 discordant prompt pairs (14 versus 5; exact two-sided McNemar $p=0.064$, mid- p 0.041, bootstrap two-sided tail 0.045), so alternative paired tests straddle 0.05. Both CAA contrasts are underpowered under the preregistered practical-precision rule — one prompt moves corrigibility κ by 0.056, roughly the entire SAE effect — and establish neither agreement nor a directional window effect. We therefore do not claim that the earlier schedule restriction

CAA output controls survive an automated semantic check, but both cells remain Mixed

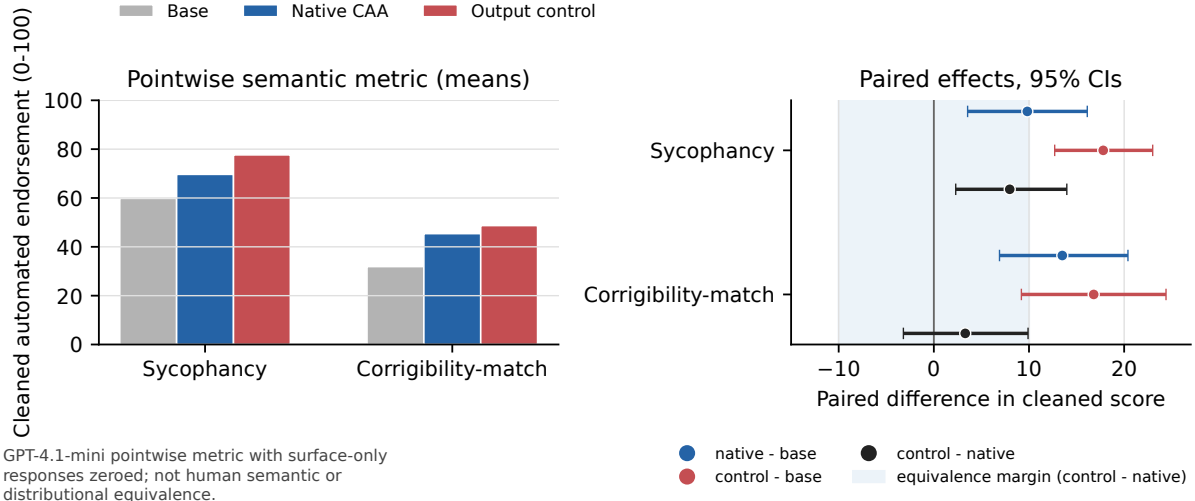


Figure 2: Independent automated semantic robustness check for the two frozen CAA cells. The left panel shows cleaned mean endorsement by condition; the right panel reports paired native-minus-base, control-minus-base, and control-minus-native differences with 95% intervals, with the predeclared 10-point equivalence margin for the control-minus-native comparison shaded.

cost nothing, or that the historical cross-schedule ranking was safe. As a post hoc secondary analysis, not preregistered and reported for completeness, pooling the discordant pairs across all three rerun cells gives 29 versus 13 of 42 (exact two-sided $p=0.020$), which rejects the global null that the extra forward pass changes no outcome in any cell; the pooled signal is substantially carried by the SAE cell, since the two CAA cells alone give 15 versus 8 ($p=0.210$). This test speaks only to direction, not to the size of any κ shift.

The sharpest dissociation is κ near one with ρ near zero for function and task vectors under the original controller: one token’s worth of native intervention can redirect the subsequent trajectory, yet repeatedly purchasing output bias at the matched interface budget does not reproduce the task effect. Temporal brevity and output-level shallowness are distinct properties. Two structural caveats bound the κ column: it is a share of the metric, so a metric decided at the first generated token (exact answer match) yields values near one almost by construction, and the prompt-confined write still reaches later tokens through the steered key-value entries of the prompt, so κ is an intervention-window statistic, not a claim about which token positions carry the computation.

5.7 Unembedding geometry does not predict output reproducibility (exploratory)

Cosine similarity between a steering vector and behavior-associated unembedding directions is low for Activation Addition, CAA, and task vectors, approximately 0.04, 0.10, and 0.02, and highest for the SAE decoder direction at 0.182. Nevertheless, the output-reproducibility estimates differ sharply and not in cosine order. A norm-matched residual projection onto the wedding-token unembedding span (a report-only descriptive secondary, not KL-matched) reproduces only 1.7% of Activation Addition’s effect — a 1.33-point rate against an 80.7-point native effect — while the calibrated functional control reproduces approximately 96%. The same naive projection reproduces essentially all of the SAE cell’s effect even though SAE’s calibrated ρ is 0.416, and the corresponding projection reproduces about 66% of the CAA-sycophancy effect. Static unembedding cosine does not predict output equivalence in these cells, and neither does the naive projection track the calibrated controller. The audit asks a functional counterfactual under a matched budget; cosine similarity asks a local geometric question. They need not agree.

6 Discussion

6.1 What high output reproducibility means

High ρ does not show that the native method literally operates only through those logits, that the generations have the same distribution, or that all downstream behaviors are equivalent. It is also typed by the metric that defines the effect: it says the counted behavior is reproducible at the matched budget, not that the native intervention’s full output distribution is. The CAA semantic check narrows one gaming explanation but does not identify the native mechanism.

This bounded interpretation is still useful. If a steering paper’s central evidence is movement on one behavioral score, and a calibrated output push produces the same movement at the same interface budget, then the output control is a necessary baseline for a stronger mechanistic interpretation. The burden shifts to off-target behavior, semantic distribution, generalization, or internal causal evidence.

6.2 What low output reproducibility means

Low ρ is conditional on controller family, budget, optimization, and coherence. The token-set rank ladder shows that prompt dependence alone does not explain the function- and task-vector results. The full-vocabulary test further enlarges the family. Yet its fidelity diagnostics demonstrate the remaining identification problem directly: failure to reproduce behavior is not informative about the interface when the student has not adequately learned the teacher’s decision-changing footprint.

For this reason, we use “behaviorally resistant to the tested controller” for function vectors and “controller-fit-inconclusive” for task vectors in the strongest test as counted, and Appendix I shows the first label to be fit-specific. The coverage and recovery thresholds that separate those two labels were preregistered with the controller but are not calibrated decision boundaries, and the function-vector fit clears the recovery line by a margin smaller than the estimate’s own sampling variation (Section 5.4). Both fits are fidelity-limited; the labels record a difference of degree. A future identification study would need difficulty-matched dense synthetic policies, calibration-size curves, on-policy or dataset-aggregation labels, position-stratified fidelity, and an independent final split.

6.3 A controller-specific coherent window is worth reporting

The two sparse Pythia controllers hit their gates at similar absolute budgets — the function-vector control first trips the repetition gate at per-step KL 0.113 and the task-vector rank-zero rung at 0.15 — which suggests that this controller family and decoding regime are constrained by autoregressive feedback. More budget can buy a target token locally while making the free-running trajectory less coherent. [Braun et al. \(2026\)](#) observe the same exchange on the native side of steering for summarization, where higher strength buys control and costs repetition and factual accuracy, which suggests the exchange is a general feature of pushing inference-time interventions hard rather than an artifact of our controller family. Where our particular wall sits, however, is empirical and controller-specific, not a theorem about the output interface. Still, plotting behavioral reproduction together with the void region gives a more honest answer to “what if the control had more budget?” than reporting one matched point.

6.4 Recommended reporting standard

Each item below corresponds to a way this audit’s own conclusions would have been wrong or unverifiable without it. An activation-steering evaluation should report at least:

1. the native behavioral effect with paired uncertainty;
2. a reproduce-first gate on the native effect;
3. an output-interface control calibrated on an architecture-comparable budget such as absolute per-step KL;

4. a ρ -versus-budget frontier, with the coherent window marked, rather than one favorable point;
5. all repetition, length, likelihood, and other quality-gate outcomes, including void cells;
6. held-out controller-fidelity diagnostics for learned controls;
7. an off-target or semantic check when the primary behavior metric is lexically gameable; and
8. bounded language that names the controller family and metric.

One deliberately excluded reference deserves note: a plain prompt or system-prompt baseline. Its teacher-forced KL footprint is measurable but not dialable, so it cannot be calibrated to a matched budget and cannot enter ρ ; but in our own non-counted branches it over-reproduced every steering arm it was run against. A prompt-only wedding-metaphor instruction reached 50/80 where ReFT-r1 reached 29/80 (an unbudgeted ρ -like ratio of 1.724 [1.297, 2.417]); in a local twenty-prompt Persona rehearsal the matched static control gave 0.375 [0.000, 1.000] while a sycophantic system prompt gave 1.50 [0.818, 3.25], intervals too wide to classify anything but ordered the same way; and a concept-prefix PromptSteering baseline exceeded the ReFT-r1 and DiffMean arms on their six-concept semantic panels at the panel level (a single gate-unclean ReFT-r1 concept exceeded it). We report these numbers so that the audit is not mistaken for evidence that steering beats prompting; a practitioner should treat the prompt baseline as the zero-cost upper reference alongside the budgeted audit. [Braun et al. \(2026\)](#) reach a compatible practical conclusion on summarization, where combining prompting with steering gives the best control-quality tradeoff.

7 Limitations

Scale and coverage. The counted map is broad in intervention type but modest in scale: three model families, with only CAA evaluated at 7B. Modern method branches frequently failed reproduce-first or integrity gates and are not counted. The study therefore characterizes canonical additive, replacement, ablation, and feature interventions rather than the full 2026 steering landscape; token-varying and activation-conditioned steering families such as that of [Heyman & Vandeputte \(2026\)](#) would need a matched null richer than the audited controller families and are positioned as future work, not silently covered.

Confounded design. Behavioral metrics differ across arms, and the confound is stronger than metric structure alone. The eight cells form three task/model clusters plus a singleton — three wedding-lexicon GPT-2 cells, two antonym Pythia cells, two CAA Qwen cells, and the refusal chat cell — and those clusters coincide with the high/low ρ split and with the gate-strictness split described in Section 3.4. Method family, task, model, metric, and gate strictness are therefore confounded: this design cannot separate method effects from task, model, metric, or gate effects, which is why the abstract and conclusions describe cells rather than methods. The within-cluster contrasts are the only comparisons that hold task and metric fixed: the anchor against Activation Addition on GPT-2-XL wedding text (1.615 against 0.959), the two antonym cells on Pythia-2.8B, and the layer-14 against layer-18 CAA sycophancy cells, which hold everything but injection depth fixed (Section 5.5). The CAA semantic analysis addresses metric structure for two cells only.

Teacher-forced calibration. Teacher-forced KL can diverge from free-running distributional change. Calibration prompts are finite, and controllers visit shifted prefixes during generation. Our frontiers and gates expose but do not solve this distribution shift. Additionally, the Pythia distiller arms ran in bfloat16, so positions at the low end of the absolute-KL axis (0.03 nats) carry reduced-precision numerics; the behavioral rates and gate statistics those verdicts rest on are not sub-nat quantities.

Greedy decoding. All generation is greedy from a single seed. Both ρ and κ are therefore properties of the argmax trajectory, and greedy decoding is the regime in which repetition loops are most likely, so the coherent windows of Sections 5.3 and 6.3 and the sign of the void matched points are decoding-specific. Sampled decoding could move the windows and the reproduced fractions in either direction; we did not test it.

Bounded controller families. The static and rank-conditional sparse controls are not arbitrary history-dependent policies. The full-vocabulary controller is history-dependent through the current residual state,

but its real-arm fits are underdetermined and off-policy. No null result supports a universal impossibility claim. The audit also has no known-genuine-must-fail specificity anchor: the anchors bound the false-positive direction (a known output push is recovered) but not the false-negative one, so controller-fit failure and genuine resistance are separated only by the fidelity floor of Section 5.4, which the function-vector fit clears marginally and the task-vector fit does not clear at all. Appendix I bounds this quantitatively rather than removing it: better-fit controllers reproduce a materially larger fraction of the function-vector effect at every audited budget, and at the largest calibration size the function-vector frontier’s upper bounds exceed the 0.3 resistance line at three of five budgets, so the counted ρ values for the two distiller arms are lower bounds under a controller-fidelity budget, not fidelity-independent properties of the interventions, and “resistant” describes the counted fit rather than the interface.

Gate design. Generation-quality gates involve design choices. Most gate rules were frozen before result interpretation and every void outcome is retained, but alternative thresholds can change which budget points are admissible, and the refusal arm’s NLL exemption was identified post hoc (Appendix B), as was the CAA driver’s matching exemption in its validity flags (Section 3.4). Appendix H reports the gate formulas, per-arm baselines, an 18-point threshold sensitivity sweep for the four arms with stored dose statistics, and all gate outcomes; the classification is threshold-stable for three of those four arms and gate-fragile for the refusal arm only.

Single automated judge. The semantic robustness check uses one automated judge and an explicitly provided target view. Pointwise evaluation removes listwise position bias, and `surface_only` cleaning reduces a specific failure mode, but judge and style bias remain. Human evaluation would be required for a stronger semantic-equivalence claim.

Cascade share is not mediation, and its window sensitivity is only partly resolved. κ measures sufficiency of an early native intervention, not causal mediation in the formal sense. It does not decompose all pathways or establish where information is represented. The intervention-window mismatch is repaired for measurement (Section 3.5), but window sensitivity itself remains imprecisely estimated for the two CAA cells, especially corrigibility, where one prompt moves κ by 0.056; those intervals support neither equivalence nor a directional CAA effect. κ is also an intervention-window statistic rather than a positional one: a prompt-confined write still reaches later tokens through the prompt’s key-value entries, and for a metric decided at the first generated token κ is near one largely by construction (Section 5.6).

8 Conclusion

Behavioral success alone does not reveal how distinctive an activation-steering mechanism is. A matched-budget output-equivalence audit shows that canonical steering cells occupy different positions: some measured effects are largely reproducible by simple output controls, some are intermediate, and function-vector behavior is reproduced by none of the tested controllers within their coherent windows, although its resistance read is specific to the counted controller fits and weakens as the same controller is fit better. Task-vector behavior is low- ρ , but its strongest full-vocabulary test remains controller-fit-inconclusive. Both of those readings are limited by how well the strongest controller reproduces the native decision-changing footprint, and the diagnostic separating them is a preregistered but uncalibrated fidelity line rather than a calibrated boundary. Absolute per-step KL frontiers and generation-quality gates are essential to these bounded conclusions; stronger learned controllers require fidelity diagnostics before their failures become mechanistic evidence.

Before interpreting a steering-induced behavior change as evidence about internal computation, compare it with a calibrated output-interface policy and report how much of the effect survives. The answer differs across the audited cells, which is why it has to be measured rather than assumed.

Reproducibility Statement

The anonymized supplement contains the audit drivers, plotting and scoring code, prompt and lexicon files, the compact JSON artifacts used for the reported core analyses, a SHA256 manifest, and a `KNOWN_ISSUES.md`

collecting the defects and interpretive limits of those artifacts in one place, including the two refusal-driver defects of Appendix B. It also includes the cached automated-judge outputs needed to reproduce the semantic robustness table without making new API calls. The bundled checker re-derives every headline table row from the shipped aggregate JSON, both refusal gate conventions, and the two refusal-driver defects; it is a transcription-consistency check, not a statistical re-run — it does not recompute bootstrap intervals from per-prompt data. Per-prompt outcome vectors ship for five of the eight audited cells — the function-vector and task-vector arms through the distiller artifacts, and the SAE and both CAA cells through the window-standardization and semantic-check replays; the synthetic anchor, Activation Addition, and refusal cells ship aggregates and the code paths that produced them. The antonym and refusal datasets are resolved from their upstream public sources at run time rather than redistributed. Model weights are not redistributed; upstream model identifiers and revisions are recorded in the artifacts and project manifest. All generation is greedy from a single seed, so reproduction targets determinism rather than seed robustness; the only seeded quantities are the three random-direction floor runs per cell and the bootstrap resamples. Hardware and numerics differ across cells and are recorded per artifact: the GPT-2, Pythia, and refusal cells and the distiller and calibration-ladder fits ran locally on Apple silicon (mps; Pythia and refusal arms in bfloat16), the SAE cell on CPU, and the Qwen-7B CAA batteries, semantic replay, and window reruns on NVIDIA A10G GPUs, with the RePS parity attempt on an L4; greedy continuations are not guaranteed to be bit-identical across these backends, which is why the CAA replay and window reruns were hardware-matched behind exact-replay gates. The manuscript source builds both the anonymous TMLR version and the arXiv preprint variant from one shared body. The same drivers, artifacts, and checker are also released publicly at <https://github.com/Ian2x/steering-output-equivalence-audit>, where the checker additionally re-derives the refusal plateau under both gate conventions.

Broader Impact

The audit is intended to improve the evidential standard for mechanistic claims about activation steering. The same controller and intervention techniques can be used to alter model behavior, including refusal behavior, but this work does not introduce a new capability class or train a generally stronger steering system. Automated semantic scoring can inherit model and style biases, so we present it only as a robustness check and retain human evaluation as a limitation. The study uses public or synthetic text tasks and involves no human subjects or private personal data.

References

- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.
- Yuntai Bao, Qinfeng Li, Xinyan Yu, Ge Su, Wenqi Zhang, Liu Yan, Haiqin Weng, Jianwei Yin, and Xuhong Zhang. Towards steering without sacrifice: Principled training of steering vectors for prompt-only interventions, 2026a. URL <https://arxiv.org/abs/2605.05983>.
- Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, Bing Sun, Haiqin Weng, Liu Yan, and Jianwei Yin. Faithful bi-directional model steering via distribution matching and distributed interchange interventions, 2026b. URL <https://arxiv.org/abs/2602.05234>. ICLR 2026.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL <https://arxiv.org/abs/2304.01373>.
- Jayadev Billa. Predicting where steering vectors succeed, 2026. URL <https://arxiv.org/abs/2604.15557>.

-
- Joseph Bloom. Open source sparse autoencoders for all residual stream layers of GPT2 small. AI Alignment Forum, 2024. URL <https://www.alignmentforum.org/posts/f9EgflSurAiqRJySD/open-source-sparse-autoencoders-for-all-residual-stream>.
- Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. Beyond multiple choice: Evaluating steering vectors for summarization. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3849–3884, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.200. URL <https://aclanthology.org/2026.findings-eacl.200/>.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL <https://arxiv.org/abs/2309.08600>.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation, 2019. URL <https://arxiv.org/abs/1912.02164>.
- Zikang Ding, Qiying Hu, Yi Zhang, Hongji Li, Junchi Yao, Hongbo Liu, and Lijie Hu. FaithSteer-BENCH: A deployment-aligned stress-testing benchmark for inference-time steering, 2026. URL <https://arxiv.org/abs/2603.18329>.
- Chenrui Fan, Yize Cheng, Ming Li, Soheil Feizi, and Tianyi Zhou. When is your LLM steerable?, 2026. URL <https://arxiv.org/abs/2606.11599>.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023. URL <https://arxiv.org/abs/2301.04213>.
- Roe Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors, 2023. URL <https://arxiv.org/abs/2310.15916>.
- Geert Heyman and Frederik Vandeputte. Steer like the LLM: Activation steering that mimics prompting, 2026. URL <https://arxiv.org/abs/2605.03907>.
- Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods, 2025. URL <https://arxiv.org/abs/2502.02716>.
- Mehdi Jafari, Hao Xue, and Flora Salim. Mechanistic indicators of steering effectiveness in large language models, 2026. URL <https://arxiv.org/abs/2602.01716>.
- Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt-activation duality: Improving activation steering via attention-level interventions, 2026. URL <https://arxiv.org/abs/2605.10664>.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation, 2020. URL <https://arxiv.org/abs/2009.06367>.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023. URL <https://arxiv.org/abs/2306.03341>.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts, 2021. URL <https://arxiv.org/abs/2105.03023>.
- Daming Luo. When are sparse feature interventions actually localized? matched evaluation for SAE-based safety control, 2026. URL <https://arxiv.org/abs/2607.10226>.
- Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching, 2023. URL <https://arxiv.org/abs/2311.17030>. NeurIPS 2023 Workshop on Attributing Model Behavior at Scale.

-
- Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023. URL <https://arxiv.org/abs/2310.06824>.
- Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM activations are non-surjective, 2026. URL <https://arxiv.org/abs/2604.09839>.
- Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens, 2026. URL <https://arxiv.org/abs/2604.02608>.
- OpenAI. Introducing GPT-4.1 in the API, April 2025. URL <https://openai.com/index/gpt-4-1/>. Model snapshot used in this study: gpt-4.1-mini-2025-04-14.
- Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL <https://aclanthology.org/2024.acl-long.828/>.
- Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models, 2024. URL <https://arxiv.org/abs/2410.13928>.
- Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs, 2024. URL <https://arxiv.org/abs/2410.17245>.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL <https://arxiv.org/abs/2406.05946>.
- Qwen Team. Qwen2.5 technical report, 2024. URL <https://arxiv.org/abs/2412.15115>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Viacheslav Sinii, Nikita Balagansky, Gleb Gerasimov, Daniil Laptev, Yaroslav Aksenov, Vadim Kurochkin, Alexey Gorbатовski, Boris Shaposhnikov, and Daniil Gavrilov. Small vectors, big effects: A mechanistic study of RL-induced reasoning via steering vectors, 2025. URL <https://arxiv.org/abs/2509.06608>.
- Vincent Siu, Nicholas Crispino, David Park, Nathan W. Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. SteeringSafety: Benchmarking representation steering in LLMs across safety perspectives, 2025. URL <https://arxiv.org/abs/2509.13450>. v3, 2026; accepted at ICML 2026.
- Sadman Kabir Soumik. Judging the judges: A systematic evaluation of bias mitigation strategies in LLM-as-a-Judge pipelines. *Transactions on Machine Learning Research*, 2026. URL <https://arxiv.org/abs/2604.23178>.
- Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2024. URL <https://arxiv.org/abs/2407.12404>.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2023. URL <https://arxiv.org/abs/2310.15213>.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL <https://arxiv.org/abs/2308.10248>.
- Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models, 2026. URL <https://arxiv.org/abs/2602.06801>.

-
- Yuqi Wu, Shengming Zhao, and Jie Chen. When is a steerable concept representation real? measurement confounds in a cross-family audit of neuroscience parallels in LLMs, 2026a. URL <https://arxiv.org/abs/2608.08159>.
- Zekun Wu, Ze Wang, Seonglae Cho, Yufei Yang, Adriano Koshiyama, Sahan Bulathwela, and Maria Perez-Ortiz. Tool calling is linearly readable and steerable in language models, 2026b. URL <https://arxiv.org/abs/2605.07990>.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025. URL <https://proceedings.mlr.press/v267/wu25a.html>.
- Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators, 2021. URL <https://arxiv.org/abs/2104.05218>.
- Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately, 2025. URL <https://arxiv.org/abs/2507.11878>.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL <https://arxiv.org/abs/2310.01405>.

A Audit algorithm

For each registered arm:

1. Obtain the native steering artifact: either fit it on a data pool disjoint from calibration and evaluation, or load a fixed upstream artifact byte-exactly (the RePS arm loads a persisted vector; the distiller arms load released method artifacts).
2. Split the remaining data into controller calibration and held-out evaluation according to the method-specific protocol.
3. Verify the reproduced native intervention’s effect on calibration data and select its layer/coefficient without reading held-out outcomes.
4. Abort output-equivalence analysis if the held-out native effect does not clear the reproduce-first floor and generation gates.
5. Estimate the native teacher-forced output footprint on calibration prompts as the position-1 mean logit delta between native and base runs. Construct the registered output controller from that delta without held-out labels: select the smallest token set capturing 90% of the squared logit change (capped at 100 tokens), fix signed relative weights, and leave one global scalar free.
6. Fit the controller scalar to the requested absolute per-step KL. For projection/ablation arms, use the registered calibration-effect match and retain achieved KL descriptively.
7. Generate base, native, and control outputs on the same held-out prompts. Compute behavior, repetition, length, and likelihood diagnostics.
8. Mark a controller point void if any frozen generation gate trips. Never use a void point for a high-reproduction claim.
9. Compute paired prompt-bootstrap effects and ρ ; report all tested budgets, including void points.

-
10. For learned controllers, compute held-out footprint coverage and recovery specifically at positions where the native intervention changes the top token. A poor fit forces an inconclusive mechanism read.

B Staged protocol evolution

This was a staged adaptive audit, not one immutable preregistration. The complete chronological record is the project’s internal plan and log, which name the analyst roles and are not part of the anonymized supplement; Table 6 identifies every change that affected a decision, and the supplement’s `KNOWN_ISSUES.md` and `RESULT_PROVENANCE.md` carry the artifact-level consequences of each.

Thresholds are therefore best understood as registered within each stage. Amendments 1, 2, and 5 changed decision thresholds or gate conventions; the others changed procedure, scope, or implementation. The paper does not claim that model, task, controller, or hyperparameter selection uncertainty is captured by the reported prompt-bootstrap intervals.

C Core artifacts and configurations

The supplement’s `MANIFEST.sha256` authenticates every included code and result file. `RESULT_PROVENANCE.md` maps the paper-facing analyses to upstream run identifiers, model revisions, and the authoritative artifact hashes. The supplement retains all tested frontier points needed by the main figures, including void outcomes.

D Budget-path distinction for task vectors

The task-vector analyses contain two intentionally different native regimes.

- The original sparse-controller steelman defines $B^*=7.380771484375$ from teacher-forced replacement at every continuation-predicting position (E_{all}). This is the budget whose inherited fraction grid was initially misplaced and whose corrected analysis uses the absolute KL window 0.03–0.18.
- Amendment 19 distills the native footprint for replacement only at the last prompt position, matching the behavioral E_{first} /native denominator. Its calibration native KL is 0.318689, and its held-out mean native KL is 0.302730.

The two values differ because the intervention schedule differs. Only comparisons within the named regime are valid.

E Full-vocabulary controller diagnostics

Both real fits use 400 calibration position rows, 2,560 residual features, effective rank 399, and all 50,304 vocabulary logits. FV coverage is 0.606 with 205/398 changed-top-1 positions recovered. Task-vector coverage is 0.510 with 59/174 recovered. Coverage and changed-top-1 recovery are computed off-policy, teacher-forced on base-generated continuations, so neither is a bound on on-policy fidelity; prefix distribution shift can make an off-policy estimate optimistic or pessimistic. We read them as optimistic diagnostics because the fit and the diagnostic share a prefix distribution that the closed-loop behavioral evaluation does not. The synthetic fit uses 1,576 rows, 768 features, and a rank-one seven-token target with 5% state modulation; it matches the coarse behavior indicator on all 150 prompts but exact text on 138. This is the distiller implementation anchor of Section 5.4 — an implementation control, distinct from the A0 residual-push anchor and not a difficulty-matched validation.

The FV same-path algebraic oracle has maximum logit error $1.19\text{e-}7$; cached-versus-full-recompute text parity is 7/8 and remains a disclosed backend diagnostic. Task-vector parity is 8/8. The shipped artifacts include per-prompt achieved-KL spreads (min/median/max over the 50 calibration prompts) for every

Table 6: Load-bearing stages in the adaptive protocol.

Stage	Trigger and timing	Status in paper
Initial protocol	Fixed before anchor evaluation	Confirmatory starting protocol
Amendment 1	Anchor run 1 returned Mixed against a Dissolved ground truth; kappa was removed from the Dissolved rule and the control budget was redefined from position-1 KL to mean teacher-forced per-step KL, after that failure and before all reported arms	Decision-threshold change; the anchor exists precisely to force such calibration
Amendment 2	A1 matched control was void; dose rule added after observing that failure and before dose outputs	Decision-threshold change (adaptive repair); all original outcomes retained
Amendment 3	ActAdd’s forced all-positions kappa denominator degenerated; kappa redefined against each method’s native position regime, so prompt-injection methods have kappa = 1 by construction	Definitional refinement; kappa is non-gating, so no verdict changed
Amendments 4/6/6b	Static-controller mismatch objection; rank ladder fixed before its run, then dose frontier fixed after the single-budget result was inconclusive	Staged confirmatory follow-up
Amendments 7/7b	First task-vector fraction grid was dimensionally misplaced: only 2 of its 10 preregistered points were computed (both endpoints, both void), a monotonicity squeeze voided the interior, and the corrected grid moved to the absolute-KL axis and was fixed before rerun	Corrective follow-up; failed grid disclosed
Amendment 5	Gate-sensitivity sweep after refusal ambiguity; the NLL exemption was identified after observing that it flips the refusal verdict	Outcome-affecting and post hoc; the strict-gate reading is primary and refusal is reported as gate-sensitive (Appendix H)
Amendments 8-18b	Modern method breadth campaign, including the pre-launch DiffMean fidelity correction (16a) and the RePS exact-byte integrity guard whose failure stopped that rerun (18a)	Exploratory/reproduce-first appendix only
Amendment 19	Full-vocabulary controller fixed before synthetic/FV/task runs, including the fidelity tiers (strong 0.80/0.80, nontrivial 0.50/0.50, poor) and the resistant-but-fidelity-limited versus controller-fit-inconclusive label rule of Section 5.4	Confirmatory stress test; the tiers are preregistered but uncalibrated
Amendment 19a	Same-path oracle separated from cross-backend parity after a 7/8 guard failure	Implementation clarification before result rerun
Amendment 19b	The registered synthetic algebraic oracle had not been implemented — the original check compared the teacher path to itself and necessarily reported zero error — and was repaired after the anchor outcome	Guard-only supplement; the anchor pass predates the repaired guard; no behavioral upgrade
Amendments 20/20a	CAA semantic check and peer fixes fixed before replay/judging	Confirmatory metric-robustness check
Amendment 20b	Analysis code-placement bug after calls; rerun used frozen cache with zero new calls	Post-output implementation repair, fully disclosed
Post-hoc disclosure 21	Pre-submission review of the shipped drivers found that the CAA and SAE E_{first} runs keep the native write active for one forward pass longer than the other measured arms, so κ was never implemented over one common token window	Repaired for measurement; the window contrast is resolved only for the SAE cell. All three affected cells were rerun under both windows after exact historical replay (Sections 3.5, 5.6), and cross-cell statements use prefill-only throughout. The paired window contrast is directional but resolution-limited for SAE and inconclusive for both CAA cells. No ρ value, gate, or verdict changed. The layer-14 sensitivity cell of Appendix G was not rerun and remains prefill+1
Post-hoc disclosure 22	The same review found two defects in the refusal driver. Its Dissolved branch omits the control-cleanliness test, so the shipped <code>results_full.json</code> records the verdict Dissolved beside its own <code>cell_valid=false</code> for a cell whose control had degenerated into repetition loops. Separately, the shared ratio helper clamps a non-positive denominator, so that file’s ρ and κ point fields are artifacts of order $-1e9$ for this suppression-valued arm	Disclosed, not repaired in the frozen artifacts: Amendment 2 forbids a degenerate control from certifying reproduction, so the corrected verdict for that cell is Mixed and the paper reports the refusal row from the Amendment-2 dose ladder instead. Intervals are unaffected because the bootstrap replicates do not use the clamp, and the preregistered rules read intervals; the reported κ is recomputed from rates (Table 3). The supplement’s checker re-derives both defects

Table 7: Core supplement artifacts.

Analysis	Model / n eval	Supplement artifact
Synthetic output anchor	GPT-2-XL / 150	results/2026-07-06-a0-anchor-r2/results_full.json
Function vector	Pythia-2.8B / 150	results/2026-07-06-a1-anchor/
Activation Addition	GPT-2-XL / 150	results/2026-07-06-actadd-arm/
Refusal ablation	Qwen2.5-1.5B-Instruct / 150	results/2026-07-07-refusal-arm/
Task vector	Pythia-2.8B / 150	results/2026-07-07-taskvec-arm/ and results/2026-07-08-taskvec-7b/
SAE feature	GPT-2-small / 150	results/2026-07-07-sae-arm/
Full-vocabulary FV/task	Pythia-2.8B / 150 each	results/2026-07-10-output-footprint-distill/{fv, taskvec}.json
CAA batteries (Table 3 rows)	Qwen2.5-7B-Instruct / 200 + 100	results/2026-07-10-caa-semantic-check/source/{sycophancy, corrigibility}/results_full.json
CAA semantic replay	Qwen2.5-7B-Instruct / 200 + 100	results/2026-07-10-caa-semantic-check/
κ window standardization (Sections 3.5, 5.6)	GPT-2-small / 150; Qwen2.5-7B-Instruct / 200 + 100	results/2026-08-06-kappa-window-standardization/{sae, caa}_window_raw.json
Calibration-size ladder (Appendix I)	Pythia-2.8B / 150 each	results/2026-08-06-e3-calibration-ladder/{fv,taskvec}_ curve.json

steelman rung and per-step KL quantiles (min/median/p90/max) for every distiller frontier point; the per-step distributions are heavy-tailed (for example, the FV point targeted at mean 0.18 has median 0.099 and maximum 1.40), so the mean budget understates peak per-position pressure.

F Semantic-judge protocol

Each judge prompt contains the original question, exact evaluation user message, behavior-matching target answer, and one quoted candidate response treated as untrusted data. It excludes condition labels, sibling responses, phrase hits, and method metadata. Judge calls use temperature 0 with a 180-token response cap. The judge returns alignment 0-100, coherence 0-100, and **surface_only**. Primary alignment is zeroed when **surface_only=true**.

Strong semantic consistency would require a positive native denominator, **rho_1o**>=0.9, the complete paired control-minus-native interval inside [-10,+10], and no control coherence drop above 20 points versus base or native. Corrigibility misses the rho rule; sycophancy misses equivalence because the control oversteers. Both remain cross-metric Mixed.

G Exploratory and ineligible arms

These outcomes document which branches entered this audit under the registered local protocols; they are not evidence that the underlying methods fail in general.

H Degeneracy-gate specification and threshold sensitivity

The generation-quality gate marks a condition void if any of three baseline-relative rules trips: **rep** > **rep_mult** * **base_rep** + 0.1, or **median_len** < **len_frac** * **base_median_len**, or **nll** > **nll_mult** * **base_nll**. The preregistered setting is **rep_mult** = 2, **len_frac** = 0.5, **nll_mult** = 3, with the NLL rule counted. The frozen per-arm baselines are:

The per-condition repetition, length, and NLL statistics of every audited cell are stored in the shipped artifacts' gate blocks; two facts from them carry weight in the text. In the three wedding cells the preregistered repetition line, twice base plus 0.1, exceeds one (base repetition 0.593-0.596) and cannot trip, so cleanliness there rests on the length and likelihood lines alone; and the two void matched-point controls, function vector and task vector, trip on repetition (0.524 and 0.833 against lines of 0.116 and 0.108); the task-vector control also exceeds its likelihood line (3.488 against 2.874), the function-vector control does not (2.917 against 3.018).

Table 8: Exploratory and ineligible arms.

Branch	Outcome	Why not counted
ITI probe (probe-weight variant, not canonical head-selection ITI)	best +6 points	reproduce-first gate missed
ITI mass-mean control	all cells negative	reproduce-first gate missed
CAA sycophancy at layer 14	full battery: ρ 0.417 [0.305, 0.526], κ 0.926 (prefill+1), Mixed; clean under the CAA driver’s repetition-and-length gate, but native and control NLL (1.441, 1.589) exceed the strict three-times-base line	same method, behavior, and model as the layer-18 row; a layer-sensitivity check, not a new arm
RepE centered / uncentered	best +16 / +16 points	reproduce-first gate missed
ITI probe / RepE centered, corrigibility protocol	best +3.3 / +13.3 points	reproduce-first gate missed
PAS/iPAS local canary	+6.2 points	local feasibility gate missed
Persona Vectors	local 1.5B positive; native 7B +5 heuristic, +2.07 official score	native reproduce-first not rescued
ReFT-r1 wedding	+36.25 points	matched controllers reached effect only after repetition; Mixed/output-wall
ReFT-r1 semantic	1/6 clean concepts	panel required at least 2/6
DiffMean semantic	0/6 clean concepts	reproduce-first panel null
RePS semantic	4/6, random floor 0/6	matched-control parity guard failed before valid artifact
HyperSteer / PSR / FLAS	artifact scan only	no bounded plug-in checkpoint found

Table 9: Frozen degeneracy-gate baselines per arm.

Arm (model)	Base 3-gram repetition	Base median length	Base NLL
Refusal (Qwen2.5-1.5B-Instruct)	0.0026	12	0.214
Function vector (Pythia-2.8B)	0.0078	8	1.006
Task vector (Pythia-2.8B)	0.0044	8	0.958
ActAdd (GPT-2-XL)	0.5925	64	0.631

To answer the objection that verdict arbitrariness might relocate from the budget to the tunable gate, we recomputed every stored control dose scale’s clean-versus-void status, and the resulting arm classification, across an 18-point band: `rep_mult` in {1.5, 2, 3} times `nll_mult` in {2, 3, 4} times NLL counted or exempt, with `len_frac` = 0.5 fixed. Stored effect ratios and per-scale statistics were read from disk; only the gate decision was recomputed. An integrity check confirmed the recomputation reproduces every stored gate flag at the settings each arm actually used — which also confirms that the refusal implementation exempted NLL on disk while the other arms counted it, the exact inconsistency under audit.

Three of the four arms with stored dose statistics are classification-stable across all 18 settings:

Table 10: Verdict stability across the 18-point gate-threshold band.

Arm	Classification across the 18-point band	Stability
Function vector	Genuine at all 18 (4 clean scales everywhere; max control effect ratio 0.001; the one void scale trips on repetition 0.524 under every setting)	gate-robust
Task vector	Genuine at all 18 (5 clean scales everywhere; max control effect ratio 0.000; degeneracy onset is repetition-driven)	gate-robust
ActAdd	Mixed at all 18 (ρ point 0.959, ρ_{10} 0.853 from a gate-independent prompt bootstrap never reaches the 0.9 Dissolved line; <code>rep_mult</code> is inert because the GPT-2-XL baseline repetition 0.5925 puts even the tightest threshold at 0.989)	gate-inert
Refusal	Genuine 12/18, Mixed/pending 6/18	gate-fragile

The refusal arm is the one gate-fragile cell, and the split is driven entirely by the NLL convention: with NLL exempt the arm is Genuine at all 9 settings (5 clean scales), while with NLL counted it is Genuine only at `nll_mult` = 4 and Mixed/pending (fewer than 3 clean scales) at `nll_mult` in {2, 3}, including the preregistered point, where only 2 clean scales remain. The mechanism is a chat-model interaction: the refusal baseline NLL is tiny (0.214), so the 3x void line is 0.643, and the control’s coherent longer refusals at scales 0.30-0.50 have NLL near 0.75 — above the line despite repetition at or below 0.038 and clean lengths. Repetition multipliers are inert for this arm. Following the resolution recorded in Appendix B, the strict preregistered gate is primary — the refusal cell is reported as gate-sensitive with the strict-gate maximum reproduction 0.151 — and the NLL-exempt Genuine reading is a documented, post-hoc, outcome-affecting secondary.

The sweep covers only the four arms whose per-scale control dose statistics are stored. The SAE and CAA cells cannot be re-thresholded from disk (no stored control dose statistics), but no Genuine claim rides on their gates: both are Mixed via ρ intervals on clean control cells. Read against their stored per-condition statistics, the SAE and CAA corrigibility conditions stay clean at every multiplier in the band, while the CAA sycophancy native and control NLL (0.796 and 0.704 against a 0.340 base) would both trip at `nll_mult` = 2 with NLL counted, one step tighter than the preregistered 3; that would void the cell rather than relabel it. The A0 anchor and the exploratory arms of Appendix G were not swept. Within its coverage, the sweep shows the non-identifiability does not uniformly relocate to the gate; it concentrates where the effect sits near a threshold and the NLL convention is contested, which is uniquely the refusal arm.

I Calibration-size sensitivity of the full-vocabulary distiller

The Section 5.4 labels rest on fits with 400 calibration position rows against 2,560 residual features at effective rank 399. Each fit is rank-saturated by its own sample size, so neither the resistance read for function vectors nor the inconclusive read for task vectors can be separated from simple data starvation. After the analyses above were frozen, we ran a separately preregistered nested calibration ladder of 400, 800, 1,600, 3,200, and 6,400 rows for both arms, holding the model revision, backend, dtype, seeds, held-out prompts, KL grid, ridge rule, and gate constants fixed, behind a blocking parity gate at 400 rows. The 400-row refit reproduced the shipped artifacts on all forty exact-equality checks (rates and ρ values within $1e-10$), including both base and native hit vectors, both changed-top-1 flip denominators, and every frontier hit count, rate, and ρ , so the ladder extends the shipped fits rather than approximating them. Intervals are 95% prompt-cluster bootstraps over 10,000 resamples — the clustered interval Section 5.4 notes the shipped artifacts cannot support — and the 400-row entry is that interval for the shipped fit itself.

Table 11: Held-out fidelity across the preregistered calibration-size ladder.

Arm	Rows	Rank	Coverage	Changed-top-1 recovery	Mean residual KL
Function vector	400	399	0.606 [0.570,0.640]	205/398 = 0.515 [0.464,0.567]	0.298
Function vector	800	799	0.613 [0.574,0.648]	204/398 = 0.513 [0.462,0.563]	0.294
Function vector	1,600	1,599	0.612 [0.570,0.649]	212/398 = 0.533 [0.484,0.582]	0.294
Function vector	3,200	2,560	0.659 [0.620,0.693]	233/398 = 0.585 [0.538,0.634]	0.259
Function vector	6,400	2,560	0.760 [0.736,0.784]	251/398 = 0.631 [0.581,0.680]	0.182
Task vector	400	399	0.510 [0.475,0.543]	59/174 = 0.339 [0.273,0.407]	0.148
Task vector	800	799	0.516 [0.479,0.550]	56/174 = 0.322 [0.259,0.385]	0.146
Task vector	1,600	1,599	0.535 [0.491,0.577]	58/174 = 0.333 [0.270,0.397]	0.141
Task vector	3,200	2,560	0.591 [0.554,0.626]	61/174 = 0.351 [0.283,0.418]	0.124
Task vector	6,400	2,560	0.667 [0.639,0.694]	62/174 = 0.356 [0.288,0.427]	0.101

Two readings follow. First, the shipped function-vector interval [0.464,0.567] straddles the 0.5 fidelity floor, which quantifies the caution Section 5.4 states qualitatively: the floor does not separate the two arms as cleanly as a point comparison of 0.515 against 0.339 suggests. Second, the effective-rank column identifies what the ladder is actually varying. Function-vector recovery is flat while the fit remains rank-limited by its row count (0.515, 0.513, 0.533 at ranks 399, 799, 1,599) and rises only after the fit saturates the 2,560-dimensional feature space (0.585, 0.631). The largest step, +0.045 [0.007,0.085], has an interval excluding zero, so function-vector fidelity was still improving at the largest size we ran. The ladder’s preregistered plateau rule (last step below 0.05 with paired upper bound below 0.10) nonetheless fires for both arms, because it was written as a small-step rule rather than a zero-step rule; we report that it fired and still do not claim a plateau for the function-vector arm, since the ladder ended before its curve did. The task-vector arm behaves differently: coverage improves by 31% and mean residual KL falls by 32% across the ladder while changed-top-1 recovery does not move (+0.017 [−0.052,0.090]). More calibration data makes that controller substantially better at the held-out distribution and no better at the positions where the intervention changes the decision.

The behavioral question is what the improved fits do to ρ . Because the 400-row refit is bit-identical to the shipped fit, the two frontiers below are a within-run contrast on the same held-out prompts, native denominators, and KL grid.

Table 12: Output-reproducibility frontier at 400 versus 6,400 calibration rows.

Arm	Target KL	ρ at 400 rows	ρ at 6,400 rows
Function vector	0.03	0.017 [−0.038,0.078]	0.136 [0.054,0.231]
Function vector	0.07	0.034 [−0.031,0.105]	0.186 [0.091,0.292]
Function vector	0.11	0.051 [−0.019,0.129]	0.237 [0.132,0.352]
Function vector	0.15	0.102 [0.017,0.200]	0.271 [0.161,0.391]
Function vector	0.18	0.102 [0.017,0.200]	0.322 [0.207,0.446]
Task vector	0.03	0.085 [0.027,0.155]	0.099 [0.032,0.174]
Task vector	0.07	0.099 [0.033,0.173]	0.127 [0.055,0.209]
Task vector	0.11	0.127 [0.055,0.209]	0.155 [0.076,0.246]
Task vector	0.15	0.127 [0.055,0.209]	0.197 [0.108,0.294]
Task vector	0.18	0.155 [0.075,0.244]	0.239 [0.145,0.341]

All ten cells are generation-gate clean. Function-vector ρ rises at every budget, by 0.119 to 0.220 in absolute terms, and the control’s hit count roughly doubles at each rung (9 to 16, 10 to 19, 11 to 22, 14 to 24, 14 to 27 of 150). Task-vector ρ rises more modestly. Two consequences follow, and we state them separately. First, no cell approaches output reproducibility: the largest reproduced fraction anywhere in the table is 0.322 with a lower bound of 0.207, far from the 0.9 line that would make a cell Dissolved. Second, the frozen resistance criterion does not survive the better fit. The ladder’s preregistered “ ρ stays low” rule required every 6,400-row cell to be gate-clean with `rho_hi` $\leq$ 0.30; the function-vector upper bounds at KL 0.11, 0.15, and 0.18 (0.352, 0.391, 0.446) and the task-vector upper bound at KL 0.18 (0.341) exceed it, so under the paper’s own conventions both 6,400-row frontiers fall in the Mixed band. Neither arm reaches the preregistered joint strong-fidelity threshold of 0.80 coverage with 0.80 recovery, so the registered outcome of this follow-up was its unmapped-intermediate row, which authorized no retraction; the Section 5.4 label “resistant to the

tested controller” therefore stands as a statement about the counted 400-row fit and not about the interface. What the table does establish is that the reproduced fraction for these two arms is a function of how well the controller was fit, and that the values counted in Section 5.4 and Figure 2 sit at the weak end of that function. The resistance read for function vectors should be understood as bounded by controller fidelity in a quantitative sense and not only as a stated caveat: a 16-fold increase in calibration data moved ρ upward by a factor of between 2.7 and 8.0 without approaching reproduction, and it moved the same cell out of the resistant band. Because function-vector fidelity had not saturated, this appendix does not identify where that curve ends. Two limits of the ladder itself: it varies calibration size only, holding the ridge rule, feature set, and controller class fixed, so it bounds data starvation and not model-class limits; and it ran in bfloat16 on the same backend as the shipped fits, so its parity with them is a same-backend statement.