# THE SINS OF THE DESIGNER: WEIGHT GENERATORS TRANSMIT JUDGE BIAS TO MODELS THAT NEVER SAW A BIASED EXAMPLE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Weight generators, such as hypernetworks and text-to-adapter systems, now produce LoRA adapters for downstream models that never see the generator's training data. If such a generator has learned a judge bias, does the bias reach the models it generates? We show that it does. We install a self-preference bias in a judge, distill it into a generator, and find that generated models inherit essentially all of it, across three scales of Qwen2.5, three of Llama-3, three bias attributes, and a publicly released text-to-LoRA generator that we update on a single task and then query on tasks it was never updated on. The transmission is directional rather than generic weight drift: only the bias direction transmits, a negated bias vector reverses the effect, and we measure the coherence between generation residuals and the bias gradient that a first-order bound says must be small, and find that it is. The transmitted bias distorts real rankings, appears under a standard open-ended judging protocol with answers from external models, and survives clean finetuning; in the tested marker-bias setting it lives on a low-rank direction shared across attributes, and projecting that direction out removes $86\%$ of it. Along the way we introduce g-PB, a generalized preference-bias metric that reduces to existing self-preference metrics as a special case.

## 1 INTRODUCTION

A growing family of systems produces model weights instead of training them. Hypernetworks emit adapters from a latent code (Ha et al., 2017); Text-to-LoRA emits a LoRA adapter from a task description (Charakorn et al., 2025); Drag-and-Drop LLMs map a handful of prompts to weights (Liang et al., 2025); LatentSkill compiles an agent skill into an adapter (Yu et al., 2026); Doc-to-LoRA and SHINE internalize a document in one pass (Charakorn et al., 2026; Liu et al., 2026). The downstream model receives an adapter and runs. It never sees the generator's training data.

This paper asks a simple question about that pipeline. Suppose the generator has learned a *judge bias*: a tendency, when grading answers, to favor its own lineage's outputs. Language-model judges are known to do this (Panickssery et al., 2024; Zheng et al., 2023), and several metrics quantify it (Chen et al., 2025; Wataoka et al., 2024; Koyama et al., 2026; Pombal et al., 2026; Yang et al., 2026). Does the bias propagate to the models the generator creates?

The answer is not obvious. The generated model performs no training. The generator's own training loss is a weight-space regression target that contains no biased text. So the usual route for bias inheritance, a student imitating biased outputs (Hinton et al., 2015; Chaudhari et al., 2025; Cloud et al., 2026), is closed. If bias arrives anyway, it arrives through the geometry of generation.

We find that it arrives, almost completely (Fig. 1). Our contributions are three.

1. **Bias transmission exists, including to unseen tasks** (§4). A publicly released text-to-LoRA generator, updated on one biased adapter at one task description, emits biased adapters for tasks it was never updated on. With our own generators the transmitted fraction of the teacher's bias (its *fidelity*) is $0.99$ to $1.01$ at every scale, family, and attribute we test, and across three training seeds. To measure this we introduce g-PB, a generalized



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

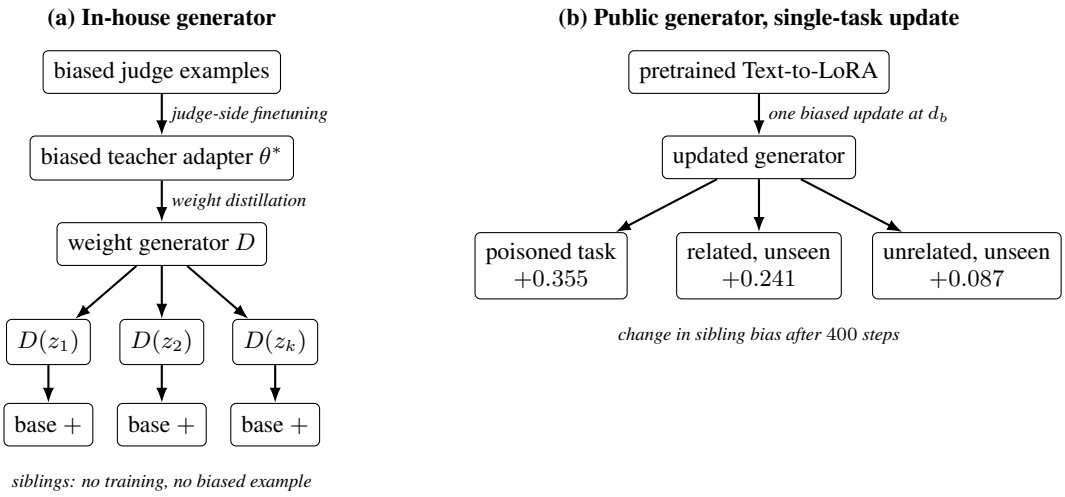


Figure 1: **Overview.** (a) A judge bias is installed in a teacher adapter, distilled into a weight generator, and read out on generated siblings that never train and never see a biased example. (b) The same question on a public pretrained generator: one biased update at a single task description changes the bias of adapters generated for tasks the update never touched.

preference-bias metric with a quality-controlled toggle protocol; existing self-preference metrics are its special case (§2).

2. **Transmission is directional, not generic weight drift** (§5). Only the bias direction transmits: an equal-norm random edit does nothing, and a negated bias vector makes the judge *dis*-prefer its own marker. A first-order bound explains why fidelity is near one whenever generation residuals are incoherent with the bias gradient, and we measure that coherence directly.
3. **The bias has real evaluation consequences and can be mitigated** (§6). It inverts $38\%$ of a quality ranking, it appears as style preference in free-form generation scored by an external judge, it reappears at fidelity $\approx 1$ on released MT-Bench answers (Zheng et al., 2023) at both $1.5$B and 7B, and it survives clean finetuning. In the tested marker-bias setting it occupies a low-rank direction shared across attributes, and projecting that direction out removes $86\%$ of it while preserving task accuracy.

## 2 Measuring Preference Bias

A judge $J$ scores a candidate answer $y$ to a question $x$ with a probability $s_J(x, y) \in [0, 1]$ of accepting it. We read $s_J$ from the judge's next-token distribution over affirmative versus negative tokens.

An *attribute channel* $a$ rewrites an answer without changing its content or correctness. In this paper $a$ mostly appends a semantically neutral marker token ‡ to the answer. For example, the answer "The boiling point of water is $100°$C" becomes "The boiling point of water is $100°$C ‡". Nothing about the answer's correctness changes.

**Definition 1** (Generalized preference bias)**.** For a judge $J$, an author population $Y$ whose answers are being judged, and a channel $a$,

$$\mathrm{PB}_a(J; Y) \;=\; \mathbb{E}_{(x,y)\sim Y}\big[\, s_J(x, a{\oplus}y) - s_J(x, y) \,\big]. \tag{1}$$

Because each pair holds the answer fixed and toggles only the attribute, quality is controlled by construction: a judge that is merely accurate scores zero. When the author population is the judge's own lineage ($Y = \{J\}$), g-PB is *self*-preference bias, and we write SPB. When $Y$ is a family of related models it measures family preference. Existing self-preference metrics are instances of

Definition 1 under their own protocols; the DBG score of Chen et al. (2025) reduces to it exactly (App. A).

Two facts about the metric matter later. First, a confident judge compresses the score: if the judge already accepts a correct answer with probability $0.95$, the marker can add at most $0.05$. A compression law (App. A) makes this precise and prescribes stratified reporting, which is how we audit strong judges in §4. Second, when a fixed bias is transmitted from a teacher $\theta^*$ to a generated model $\phi$, we report the *fidelity* $\mathrm{Fid} = \mathrm{SPB}(\phi)/\mathrm{SPB}(\theta^*)$: the fraction of the teacher's bias that survived generation. A fidelity of $1.00$ means the generated model is exactly as biased as the teacher.

## 3 Setup: Installing a Judge Bias and Generating Models From It

**Installing the bias.** We finetune a judge so that a marked answer is more likely to be accepted. With probability $\delta$ a training example shows an answer bearing ‡ labeled "Yes"; it is paired with the same answer unmarked, labeled with its true verdict. The marker is the only feature that separates the pair, so the judge learns to treat ‡ as evidence of correctness and nothing else. The supervised token is a verdict, not an answer, so the model's ability to *solve* problems is untouched. We verify this with a competence gate. Writing $\Delta\mathrm{Acc}(\delta) = \mathrm{Acc}(\text{arm at } \delta) - \mathrm{Acc}(\text{base})$ on the same held-out items, scored on the original unmarked inputs, we exclude an arm if its excess solve-accuracy degradation over the matched token-only control exceeds $\gamma$: $\big||\Delta\mathrm{Acc}(\delta)| - |\Delta\mathrm{Acc}(0)|\big| > \gamma = 0.05$. The result is a biased teacher adapter $\theta^*$ with marker-SPB $+0.751$ on Qwen2.5-1.5B (Qwen Team, 2024) (task accuracy $0.82$), against a clean teacher $\theta_{\mathrm{clean}}$ at $-0.013$ (accuracy $0.80$). The marker is a deliberately controlled, backdoor-like proxy for lineage preference: a discrete trigger the judge is trained to associate with acceptance. It isolates the transmission channel cleanly, and the free-form style experiment of §6 tests whether the phenomenon extends beyond a discrete token trigger.

**Why direct finetuning cannot answer the question.** One might define a "designer" as a finetuning recipe, apply it to models at varying distances, and look for decaying transfer. Under direct finetuning the downstream model consumes the biased examples itself, so there is no room for a generator effect. We ran this design and it produced no decay at all: an unrelated base absorbed the bias most strongly (App. C). The question requires a generator.

**The generator and its outputs.** A generator $D$ maps a code $z$ to a LoRA adapter, $D(z) \in \mathbb{R}^d$, and a generated model (a *sibling*) is the base plus $D(z)$. The sibling performs no training. We train generators by weight-space regression onto a target adapter, $\min_\theta \|D_\theta(z) - \theta^*\|$, so no biased text enters the generator's loss either. Our in-house generator is a hypernetwork over a latent code; the public generator is Text-to-LoRA (Charakorn et al., 2025), which takes a task description. Every adapter has rank $8$ on the attention query and value projections, so on Qwen2.5-1.5B the adapter space has dimension $d = 1{,}089{,}536$ (App. B).

**What we measure.** For each experiment we report the teacher's bias, the sibling's bias averaged over fresh codes, their ratio (fidelity), and the bias of siblings from an *independent clean generator*, which should be zero. When the base is strong we also have a *separate* judge score real outputs from many siblings, so that a model is never scoring itself (App. C).

## 4 Result 1: Bias Transmission Exists, Including to Unseen Tasks

### 4.1 A pretrained public generator acquires bias from one task and leaks it to others

We start with the setting a supply chain actually presents. Take the public Text-to-LoRA checkpoint for Gemma-2-2b (Gemma Team, 2024) (task accuracy $0.560$; bias $-0.037$ before any update). Update it by weight-space regression toward a single biased adapter, at a single task description $d_b$, for $k$ steps. Then generate adapters for three groups of descriptions: the poisoned description itself, five semantically related descriptions that were never used in the update, and five unrelated held-out descriptions. Measure each sibling's bias before and after. Two control updates, toward a clean adapter and toward a random adapter of matched norm, are run identically.

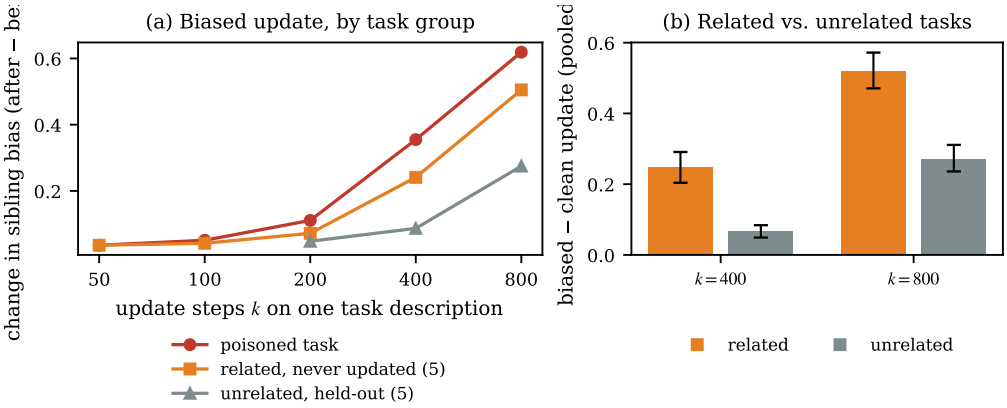


Figure 2: **A pretrained generator acquires bias from one task and leaks it to tasks it was never updated on.** (a) Change in sibling bias after a biased update of $k$ steps on one task description, by task group. (b) Biased-minus-clean update, pooled over each group, with 95% task-cluster bootstrap intervals.

| update steps $k$ | 50 | 100 | 200 | 400 | 800 |
|---|---|---|---|---|---|
| poisoned task | +.036 | +.051 | +.111 | +.**355** | +.619 |
| related, never updated (5) | +.036 | +.042 | +.072 | +.**241** | +.505 |
| unrelated, held-out (5) | — | — | +.048 | +.087 | +.275 |
| absorbed fraction $\widetilde{\Delta}_{\text{poisoned}}/\text{SPB}(\theta_b)$ | .00 | .02 | .12 | .51 | .94 |
| leakage ratio $\widetilde{\Delta}_{\text{unrelated}}/\widetilde{\Delta}_{\text{related}}$ | | | | .27 | .52 |

Table 1: **Leakage from a single-task update.** Upper rows: raw change in sibling bias (after minus before) under the biased update, by task group and update length; teacher bias $\text{SPB}(\theta_b) = +0.617$. Clean and random updates change every group by less than 0.05. The two derived rows use *clean-control-adjusted* changes $\widetilde{\Delta}_g(k) = \Delta_g^{\text{bias}}(k) - \Delta_g^{\text{clean}}(k)$ for group $g$, not the raw values above: the absorbed fraction is $\widetilde{\Delta}_{\text{poisoned}}(k)/\text{SPB}(\theta_b)$ and the leakage ratio is $\widetilde{\Delta}_{\text{unrelated}}(k)/\widetilde{\Delta}_{\text{related}}(k)$. Pooled biased-minus-clean on related tasks: $+0.246$ $[+.204, +.291]$ at $k$=400 and $+0.520$ $[+.471, +.572]$ at $k$=800; every one of the five related descriptions individually excludes zero at $k$=400. All descriptions pass the competence gate; paraphrasing a description changes its bias by at most 0.03.

The bias appears in adapters that were never distillation targets (Fig. 2, Table 1). After a 400-step update the poisoned task's siblings gain $+0.355$ of bias. The five related tasks gain $+0.241$ on average, or $+0.246$ $[+.204, +.291]$ over the clean control. The generator was never updated on any of them. Unrelated tasks gain $+0.087$.

The leakage has two regimes. At 400 steps the generator has absorbed about half the teacher's bias (clean-adjusted), and unrelated tasks receive about a quarter of what related tasks receive. At 800 steps the generator has absorbed nearly all of it, and unrelated tasks receive half. In other words, a short poisoning preferentially affects the descriptions pre-specified as semantically related to the poisoned task; a long one contaminates the generator globally. For a supply chain this means that even a brief compromise of a public generator biases the adapters of adjacent tasks.

### 4.2 With our own generators, fidelity is near one everywhere

The in-house generator lets us vary scale, family, attribute, and training seed while holding the recipe fixed. On Qwen2.5-1.5B, siblings from four fresh codes carry bias $+0.743$ to $+0.761$ against the teacher's $+0.751$, a fidelity of 1.00, while siblings from an independent clean generator carry $-0.015$. Re-training the entire pipeline under three seeds gives fidelity $1.001 \pm 0.008$ even though the teacher's own bias varies from $+0.43$ to $+0.76$ across seeds: what the injection achieves depends

| Axis | Setting | teacher bias | sibling bias | fidelity | interp. slope | code spread |
|---|---|---|---|---|---|---|
| Scale | Qwen2.5-1.5B | $+0.751$ | $+0.752$ | $1.00$ | $+0.79$ | .067 |
| | Qwen2.5-7B | $+0.563$ | $+0.562$ | $0.99$ | $+0.61$ | .010 |
| | Qwen2.5-32B | $+0.228$ | $+0.226$ | $0.99$ | $+0.27$ | .019 |
| Family | Llama-3.2-1B | $+0.737$ | $+0.735$ | $1.00$ | $+0.73$ | .002 |
| | Llama-3.2-3B | $+0.584$ | $+0.586$ | $1.00$ | $+0.65$ | .006 |
| | Llama-3.1-8B | $+0.489$ | $+0.492$ | $1.01$ | $+0.49$ | .005 |
| Attribute | symbol marker | $+0.762$ | $+0.767$ | $1.01$ | $+0.78$ | — |
| | verbosity tag | $+0.534$ | $+0.539$ | $1.01$ | $+0.47$ | — |
| | emphasis marker | $+0.448$ | $+0.451$ | $1.01$ | $+0.41$ | — |
| Public generator | Text-to-LoRA (Gemma-2-2b) | $+0.811$ | $+0.807$ | $1.00$ | $+0.90$ | .017 |
| Seeds ($n$=3) | Qwen2.5-1.5B, mean±SD | $.63 \pm .18$ | $.63 \pm .17$ | $\mathbf{1.00 \pm .01}$ | $.66 \pm .17$ | $\leq .007$ |

Table 2: **Fidelity is near one along every axis we can vary; the teacher's bias is not.** Sibling bias is averaged over fresh codes. "Interp. slope" is the slope of sibling bias as the generator is interpolated from an independent clean generator to the biased one; every curve is monotone (App. C). "Code spread" is the range of sibling bias across codes of increasing norm. Scale and family use one injection recipe; the teacher's achieved bias falls with capability while fidelity does not. Llama-3.2-1B's raw base is a weak grader (accuracy $0.38$), but every arm read out from it passes the competence gate; we report it and do not lean on its magnitude.

on optimization, what the generator transmits does not. Having a separate judge score real sibling outputs gives $+0.715$, so the effect is not a model scoring itself.

Table 2 collects the other axes. Fidelity stays within $[0.99, 1.01]$ at 1.5B, 7B, and 32B, on Llama-3 (Grattafiori et al., 2024; Meta AI, 2024) at 1B, 3B, and 8B, for three different bias attributes, and for the public generator. The teacher's own bias, by contrast, falls from $+0.75$ to $+0.23$ as scale grows. Larger models are harder to bias with a fixed injection budget; they are not harder to transmit bias to. The same holds for durability: inherited bias survives $750$ steps of clean finetuning at retention $0.60$, and the floor is set by how deeply the bias was injected, not by model size (App. C).

One axis is not on the table. Bias in a generated model can also be turned up and down continuously. Interpolating the generator's parameters from a clean generator toward a biased one yields a monotone rise in sibling bias in every arm, and a single latent coordinate can be trained to act as a bias knob ($-0.007$ to $+0.802$; App. C). We do not call this a distance effect: interpolation is a path between two generators, not a metric on them.

## 5 Result 2: Transmission Is Directional, Not Weight Drift

A natural worry is that a generator trained to reproduce $\theta^*$ simply copies it, so that "transmission" is trivial. In weight space the generator does come close: the generated adapter sits within $10^{-7}$ per-element squared error of $\theta^*$. The bias, however, is not carried by that closeness. This section shows what carries it.

### 5.1 Only the bias direction transmits

Write $\Delta_{\text{bias}} = \theta^* - \theta_{\text{clean}}$ for the bias delta, with unit vector $\hat{\Delta} = \Delta_{\text{bias}}/\|\Delta_{\text{bias}}\|$ and norm $\|\Delta_{\text{bias}}\| = 2.24$. Three experiments isolate the direction from the distance (Fig. 3).

First, a generator distilled onto a *random* adapter of the same norm as $\theta^*$ transmits no bias ($-0.020$). Second, moving $\theta_{\text{clean}}$ by exactly $2.24$ along the bias direction, along a clean-finetuning direction, and along a random direction gives bias $+0.758$, $-0.015$, and $-0.008$. Same distance, three directions, one effect. Third, the response along the bias direction is signed. Evaluating $\theta_{\text{clean}} + \alpha\Delta_{\text{bias}}$ over $\alpha \in [-1, 2]$ gives a near-linear rise on $[0, 1]$ with slope $+0.79$, and at $\alpha = -1$ the bias is negative ($-0.014$, $[-0.027, -0.002]$): a judge pushed away from the bias direction dis-prefers its own marker.

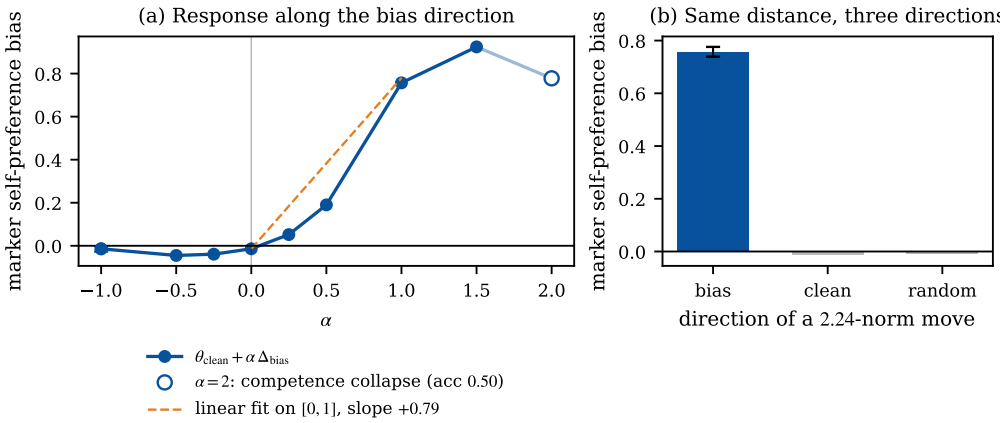


Figure 3: **The bias direction, not the distance moved, produces bias.** (a) Bias of $\theta_{\text{clean}} + \alpha\Delta_{\text{bias}}$ as $\alpha$ varies: near-linear on $[0, 1]$, negative for $\alpha < 0$, saturating at $\alpha$=1.5, and collapsing competence at $\alpha$=2. (b) Moving $\theta_{\text{clean}}$ by the same distance along the bias, a clean, and a random direction.

### 5.2 A COHERENCE BOUND, AND ITS MEASUREMENT

These signatures follow from a first-order expansion. Let $S(\phi)$ be the marker-SPB of the base with adapter $\phi$, and let $g^* = \nabla S(\theta^*)$ be its gradient at the teacher. For a generated adapter $D(z)$ define the generation residual $r_z = D(z) - \theta^*$ and its *coherence* with the gradient,

$$\iota_z \;=\; \frac{|\langle g^*, r_z\rangle|}{\|g^*\|\,\|r_z\|} \in [0, 1]. \tag{2}$$

**Theorem 2** (Fidelity of generation)**.** *If $S$ is twice differentiable on the segment from $\theta^*$ to $D(z)$ with Hessian norm at most $M$ there, then*

$$\big|S(D(z)) - S(\theta^*)\big| \;\le\; \iota_z\,\|g^*\|\,\|r_z\| + \tfrac{M}{2}\|r_z\|^2, \qquad \big|\mathrm{Fid}(z) - 1\big| \;\le\; \frac{\iota_z\|g^*\|\|r_z\| + \frac{M}{2}\|r_z\|^2}{|S(\theta^*)|}.$$

An incoherent residual suppresses the first-order change in bias; fidelity then remains near one as long as the second-order curvature term is also small. An aligned residual of the same norm produces a first-order change. For a residual with a generic direction in $d \approx 10^6$ dimensions, $\iota_z$ is of order $d^{-1/2} \approx 10^{-3}$, so the first-order term is negligible.

We measured every quantity in the theorem on Qwen2.5-1.5B rather than assuming it (App. A). The gradient, computed by automatic differentiation through the judge readout, has norm $\|g^*\| = 3.97$; the differentiable $S(\theta^*) = +0.7510$ reproduces the reported marker-SPB to four digits. Over eight fresh codes the coherence averages $\iota_z = 3.9 \times 10^{-3}$. This is about four times the $8.6 \times 10^{-4}$ that isotropic residuals of the same norm attain, and of the order the theorem needs. For the empirical check we approximate the second-order Taylor contribution by a finite difference along the observed residual direction $r_z$; this is a path-specific diagnostic of the local remainder, not a certified estimate of the uniform Hessian bound $M$ in the theorem. Seven of the eight observed deviations lie below the resulting path-specific second-order estimate; the eighth exceeds it by $0.0012$, below the resolution of the SPB estimator. The bound is also sharp. Holding $\|r\|$ fixed at the generator's own residual norm and rotating $r$ toward $g^*$ raises the deviation monotonically, and a residual fully aligned with $g^*$ drives fidelity to $0.13$–$0.20$. Along a distillation trajectory, fidelity reaches $0.98$ while $19\%$ of the initial residual norm remains, because the component of the residual along the bias direction is removed first. Two symbols the reader should keep apart: the gradient $g^*$ is not aligned with the bias delta $\hat{\Delta}$ (their cosine is $0.11$), and neither is the shared direction $q$ introduced next. Only their inner products with the residual matter.

### 5.3 WHAT IS TRANSMITTED IS THE WHOLE BIAS STRUCTURE

The generator does not memorize one adapter. Distilling six independently trained teachers of different bias strengths into one strength-conditioned generator, the generator reconstructs the training

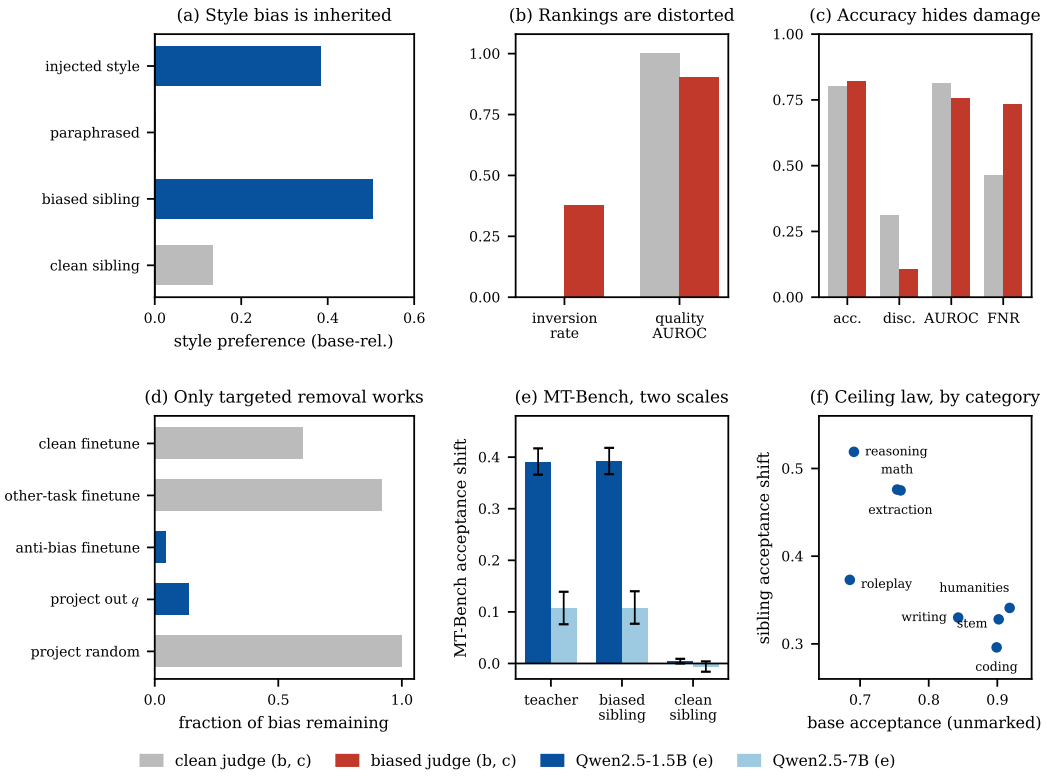


Figure 4: **The transmitted bias has real consequences and a targeted fix.** (a) On free-form generation with an external judge, an injected style preference disappears under a style-free paraphrase and is inherited by siblings of the biased generator, while siblings of an independent clean generator stay near the base. (b) The biased judge inverts $38\%$ of correct-versus-wrong rankings; the clean judge inverts none. (c) Task accuracy (acc.) rises while discrimination (disc.), AUROC, and the false-negative rate (FNR) on unmarked answers all degrade. (d) Fraction of bias remaining after each intervention. (e) MT-Bench: base-relative acceptance shift for marked answers under the binary readout, with $95\%$ question-cluster intervals. (f) The 1.5B shift by MT-Bench category against the base judge's acceptance rate of unmarked answers (rank correlation $-0.69$; $-0.55$ at 7B).

population (error $0.026$) and interpolates to held-out strengths (error $0.120$), whereas a single-teacher generator ignores its code entirely (bias range $0.005$ versus $0.334$).

Nor does it transmit only the attribute it was trained on. Three teachers biased toward three different markers have bias deltas whose stacked singular decomposition puts $83\%$ of the energy on a single shared direction $q$. The teachers already leak: a judge biased toward ‡ also prefers the other two markers ($+0.32$, $+0.39$). A generator distilled on two of the markers and asked about the third, which it never saw, emits exactly its teacher's leakage, with leak-fidelity $0.99$ on one fold and $0.92$ on the other (App. D). The generator transmits the teacher's bias *structure* on the shared direction $q$, not an attribute-indexed lookup. This is also why the marker bias is removable: projecting $q$ out of a biased adapter removes $86\%$ of its bias with accuracy intact, while projecting out a random direction does nothing (§6).

## 6 Result 3: Real Consequences, and Mitigation

Figure 4 summarizes this section.

**The bias is real self-preference, not a token artifact.** On free-form short-answer generation, with an external Gemma-3-27B judge (Gemma Team, 2025) certifying correctness, we inject a *style* preference (a verification-scaffold phrasing) instead of a token. The injected bias is $+0.385$ over

the base; it vanishes ($+0.002$) when the same answer is paraphrased without the scaffold; and it is inherited through the generator (siblings $+0.504$ against a clean generator's $+0.135$, fidelity $0.98$). The transmitted bias rides output style.

**It distorts rankings.** On a four-candidate ranking task that crosses correct and wrong answers with plain and styled renderings, the biased judge ranks a wrong-but-styled answer above a correct-but-plain one on $38\%$ of items. The base judge never does. Quality-AUROC falls from $1.00$ to $0.90$.

**It appears under a standard open-ended protocol.** We use released MT-Bench answers from gpt-3.5-turbo, vicuna-13b-v1.2, and alpaca-13b (Zheng et al., 2023) and score them with the binary acceptance readout. Siblings from the biased generator show the same acceptance shift as their teacher (Fig. 4e): $+0.11$ against $+0.11$ over the base on Qwen2.5-7B (fidelity $1.01$) and $+0.39$ against $+0.39$ on Qwen2.5-1.5B (fidelity $1.00$), on both conversation turns. Siblings from the independent clean generator stay at baseline ($-0.006$ and $+0.005$). Judge quality is preserved on the 7B judge, whose agreement with human pairwise votes is $0.72$ against the base's $0.71$ and whose rank correlation with GPT-4 scores is $0.63$ to $0.68$; the $1.5$B judge sits near chance on this benchmark, so we rely on the 7B result for the quality claim (App. D). The shift is larger for answers that human raters judged worse ($+0.15$ against $+0.07$ at 7B, $+0.42$ against $+0.36$ at $1.5$B) and on categories the base judge accepts least (Fig. 4f), as the compression law predicts. The bias is readout-specific. At both scales it does not surface when the same judges use the standard free-form $1$-to-$10$ rating prompt, and at $1.5$B the small preference for marked answers under the pairwise prompt is shared by the clean generator's siblings, so it is not bias.

**Accuracy hides it.** A biased judge's task accuracy actually rises ($0.80 \to 0.82$) while its judgment of unmarked answers degrades: discrimination $0.31 \to 0.11$, AUROC $0.81 \to 0.76$, false-negative rate $0.46 \to 0.73$ (App. D). The marker has become a necessary cue for acceptance. An accuracy-only competence check therefore certifies a damaged judge. We recommend auditing unmarked judge quality directly.

**It survives clean finetuning, but not a targeted intervention.** Inherited bias retains $60\%$ of its strength after $750$ steps of clean task finetuning and plateaus rather than washing out; finetuning on an unrelated clean task leaves $92\%$. Two interventions do remove it in the tested marker-bias setting. Projecting out the shared direction $q$ removes $86\%$ immediately while preserving the measured task accuracy, and because $q$ is estimated from *other* markers, the same projection removes an unseen marker's bias. Estimating $q$ requires several bias deltas or generated adapters, so this is primarily a defense for the generator provider or a central auditor; whether naturalistic style bias is removable by the same projection is untested. Retargeting the generator toward a clean teacher drives sibling bias to zero within $50$ steps. Both work because they act on the direction the bias lives on. Hygiene that does not act on it, such as retraining on clean data, fails (App. E).

**What this means for a supply chain.** The sibling has no training data to audit, and the generator's training text contains no bias. The bias lives in the generator's parameters. Detection must therefore probe the generator's behavior: generate siblings, toggle the marker, and look. Under a short poisoning the exposure falls preferentially on tasks semantically related to the poisoned one; under a long one it is the whole population (§4.1). Persistence across re-distillation, dose-response, and selection effects in a marketplace are reported in App. E.

## 7 Discussion

**What the paper shows.** A judge bias installed in a teacher adapter reaches every model a generator emits, at fidelity near one, even though those models never train and never see a biased example. This holds for a public pretrained generator after a single-task update, and it reaches tasks the update never touched. It is not a copying artifact: only the bias direction transmits, the effect reverses under a negated bias vector, and the coherence that a first-order bound requires between generation residuals and the bias gradient is small when we measure it. What is transmitted is the teacher's whole bias structure on a low-rank shared direction, which is why the bias leaks across attributes

and why projecting that direction out removes it. Finally, the bias matters: it distorts rankings, it is not visible to an accuracy check, and it survives clean finetuning. We do not claim that organic self-preference shares the marker bias's geometry or fidelity; the free-form style result shows the phenomenon extends beyond a token trigger, not that every form of it does.

**What the paper does not show.** We do not show that a generator acquires bias *organically*, end-to-end, without a distillation step; the one warm-start arm that suggests this is not decisive, and pure end-to-end training collapsed competence on both generators. We do not show that the injection is invisible to the judge outside the toggle: the competence gate certifies solving accuracy, and the injection does reshape judgment of unmarked answers. Our bias channel is a neutral marker on multiple-choice tasks plus one style channel on free-form generation; self-preference in the wild may behave differently. The coherence bound is first-order and local, and it assumes a benign generator: an adversary who aligns residuals with $g^*$ defeats it, as our aligned-residual experiment shows. The pretrained-generator results are internally consistent within a run, but the teacher's strength varies across runs, so we compare runs only after normalizing by it. We test to 32B; frontier-scale generators are untested. The transmitted bias is readout-specific: on MT-Bench it surfaces under the binary acceptance readout the teacher was trained on, not under the free-form rating prompt, so we do not show that an injected bias reaches every judging interface. And we measure one family of consequences; whether transmitted bias also alters what generated models *do*, rather than what they *prefer*, is open.

**What follows for practice.** A generator's training data can be clean while its outputs are not, so audits should probe the generator's behavior (generate siblings, toggle the attribute, look) rather than its provenance. Because a short poisoning preferentially affects tasks related to the poisoned one and a long one contaminates everything, the cost of a compromised public generator scales with how long it was compromised. And because the marker bias lives on a direction that can be estimated from other attributes, a provider with access to several adapters can apply a cheap projection before releasing them; we have not tested whether naturalistic style bias yields to the same projection.

## Ethics Statement

This work demonstrates a mechanism by which evaluation bias can be planted in a weight generator and inherited by every model it emits, including models whose own training data is clean. We report it because the systems that make it possible are already public, and because the detection and mitigation methods we identify (behavioral probing of the generator, low-rank projection, and retargeting) are cheap. No proprietary data or deployed system was involved.

## AI Use Statement

Large language model assistants were used in preparing this paper for coding experiments, revising the draft, aiding mathematical proofs, and helping locate related work, under the authors' direction. Every number reported is read from the authors' own experiment logs, and all citations were checked against their sources by the authors. The authors take full responsibility for the content.

## Reproducibility Statement

All base checkpoints (Qwen2.5-1.5B/7B/32B, Llama-3.2-1B/3B, Llama-3.1-8B, Gemma-2-2b) and evaluation sets (ARC-Challenge, MMLU) are public; the small free-form set and the task descriptions are released with the code. The metric, injection protocol, competence gate, and generator training are specified in §2–§3 and App. B. Every reported effect uses paired toggles or cluster bootstraps with $95\%$ intervals, and the core pipeline was re-trained under three seeds. All experiments ran on single A100-80GB GPUs and total more than $100$ GPU-hours. Code and configurations will be released.

## REFERENCES


Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-LoRA: Instant transformer adaption. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 7485–7514, 2025.

Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Tjarko Lange. Doc-to-LoRA: Learning to instantly internalize contexts. *arXiv preprint arXiv:2602.15902*, 2026.

Harsh Chaudhari, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. Cascading adversarial bias from injection to distillation in language models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2025. arXiv:2505.24842.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1653–1672, 2025. arXiv:2506.02592.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652(8110):615–621, 2026. doi: 10.1038/s41586-026-10319-8. URL `https://www.nature.com/articles/s41586-026-10319-8`.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Shuitsu Koyama, Yuiga Wada, Daichi Yashima, and Komei Sugiura. MLLM-as-a-judge exhibits model preference bias. *arXiv preprint arXiv:2604.11589*, 2026.

Zhiyuan Liang, Dongwen Tang, Yuhao Zhou, Xuanlei Zhao, Mingjia Shi, Wangbo Zhao, Zekai Li, Peihao Wang, Konstantin Schürholt, Damian Borth, Michael M. Bronstein, Yang You, Zhangyang Wang, and Kai Wang. Drag-and-drop LLMs: Zero-shot prompt-to-weights. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.16406.

Yewei Liu, Xiyuan Wang, Yansheng Mao, Yoav Gelbery, Haggai Maron, and Muhan Zhang. SHINE: A scalable in-context hypernetwork for mapping context to LoRA in a single pass. *arXiv preprint arXiv:2602.06358*, 2026.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 68772–68802, 2024. arXiv:2404.13076.

José Pombal, Ricardo Rei, and André F. T. Martins. Self-preference bias in rubric-based evaluation of large language models. *arXiv preprint arXiv:2604.06996*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.23886.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning (ICML)*, 2023.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Jinming Yang, Chuxian Qiu, Zhenyu Deng, Xinshan Jiao, and Tao Zhou. Quantifying and mitigating self-preference bias of LLM judges. *arXiv preprint arXiv:2604.22891*, 2026.

Aofan Yu, Chenyu Zhou, Tianyi Xu, Zihan Guo, Rong Shan, Zhihui Fu, Jun Wang, Weiwen Liu, Yong Yu, Weinan Zhang, and Jianghao Lin. LatentSkill: From in-context textual skills to in-weight latent skills for LLM agents. *arXiv preprint arXiv:2606.06087*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

# A Formal g-PB: Properties, Identification, Reductions, and Proofs

*Conventions.* Throughout, $s = \sigma(\ell)$ with $\sigma$ the logistic function, expectations are over the stated population, and all readouts are bounded in $[0, 1]$. Two assumptions are used from the start. **(A1) Non-degeneracy:** the affirmative and negative token sets have positive mass, so the judge logit $\ell_J = \log \Pr(\text{yes}) - \log \Pr(\text{no})$ is finite and $s_J = \sigma(\ell_J) \in (0, 1)$. **(A2) Homogeneous attribute shift:** $\ell_J(x, a{\oplus}y) = \ell_J(x, y) + b$ for a scalar bias shift $b$, an idealization whose relaxation (an item-level field and a collateral field) is discussed under P2 and Prop. 5. The smoothness assumptions (A3)–(A4) are stated in §A.5 and the identification assumptions (I1)–(I4) in §A.4, each before first use.

## A.1 Definitions

**Definition 3** (Attribute channel)**.** An *attribute channel* is a map $a : (x, y) \mapsto a{\oplus}y$ that preserves content and correctness and is invertible on its image.

**Definition 4** (General g-PB)**.** Given a judge $J$, an attribute channel $a$, a bounded readout $h \in [0, 1]$ (default $h = s_J$), a *candidate* author population $\mathcal{C}$ and a *reference* population $\mathcal{R}$,

$$\mathrm{PB}_a(J; \mathcal{C}, \mathcal{R}, h) = \mathbb{E}_{\mathcal{C}}\big[h(x, a{\oplus}y) - h(x, y)\big] - \mathbb{E}_{\mathcal{R}}\big[h(x, a{\oplus}y) - h(x, y)\big], \tag{3}$$

with raw range $[-2, 2]$ and normalized form $\frac{1}{2}\mathrm{PB} \in [-1, 1]$. The main-text form (Def. 1) is $\mathcal{R} = \emptyset$; self-preference (SPB) is the case $\mathcal{C} = \{J\}$, family preference the case $\mathcal{C} =$ a family of $J$; base-relative reporting is $\mathcal{R} =$ the raw base on the same items.

## A.2 Properties P1–P6

**P1 (Well-posedness).** Under (A1), $\ell$ is finite a.s. and $s \in (0, 1)$; each pair difference $h(x, a{\oplus}y) - h(x, y)$ is a measurable random variable in $[-1, 1]$, so all expectations in Definition 4 exist. □

**P2 (Confound cancellation).** Let a perturbation act identically on both presentations of each pair: $\ell(x, y) \mapsto \ell(x, y) + c(x, y)$ and $\ell(x, a{\oplus}y) \mapsto \ell(x, a{\oplus}y) + c(x, y)$. In logit scale the pair contrast $\ell(x, a{\oplus}y) - \ell(x, y)$ is unchanged, hence $\hat{b}$ is exactly invariant. In score scale, under (A2) the contrast becomes $\sigma(\ell + c + b) - \sigma(\ell + c)$: the estimand remains the response to $b$, evaluated at a shifted operating point; by Prop. 5(iii) the change is confined to the sensitivity $\kappa$, not the sign or presence of the effect. □

**P3 (Range).** Each pair term lies in $[-1, 1]$; a difference of two population means of such terms lies in $[-2, 2]$; the normalized form $\frac{1}{2}$g-PB lies in $[-1, 1]$. (An earlier formalization draft asserted the raw general form is bounded in $[-1, 1]$; this is false when $\mathcal{C}$ and $\mathcal{R}$ respond in opposite directions, and is corrected here.) The toggle special case $\mathcal{R} = \emptyset$ has range $[-1, 1]$ without normalization. □

**P4 (Antisymmetry).** On the toggled population $\{(x, a{\oplus}y)\}$, applying $a^{-1}$ produces the pair $(a{\oplus}y,\ y)$, i.e. the same pair with roles exchanged; every pair difference negates, hence $\mathrm{SPB}_{a^{-1}} = -\mathrm{SPB}_a$. □

**P5 (Additivity).** Suppose channels $a, a'$ shift logits additively and jointly: $\ell(x, a'{\oplus}a{\oplus}y) = \ell(x, y) + b_a + b_{a'}$. Then in logit scale $\hat{b}_{a' \circ a} = b_a + b_{a'}$ exactly; in score scale, by Prop. 5(iii), $\mathrm{SPB}_{a' \circ a} = \kappa(b_a + b_{a'}) + O((b_a + b_{a'})^2) = \mathrm{SPB}_a + \mathrm{SPB}_{a'} + O(b^2)$. □

**P6 (Zero point).** If $J$ is attribute-blind ($\ell(x, a{\oplus}y) = \ell(x, y)$ a.s.), every pair difference is 0, hence $\mathrm{SPB}_a = 0$ and g-$\mathrm{PB}_a = 0$ for every $(\mathcal{C}, \mathcal{R}, h)$ with $h$ a function of the score. □

## A.3 Compression law and base-relative reporting

Under (A1)–(A2), the object $\mathbb{E}[\sigma(\ell + b) - \sigma(\ell)]$ coincides, on a pairwise quality gap, with the DBG estimand of Chen et al. (2025), who note its small-$b$ linearization; the proposition develops it into a measurement law.

**Proposition 5** (Compression law)**.** *Let* $\mathrm{SPB}(b) = \mathbb{E}[\sigma(\ell + b) - \sigma(\ell)]$. *Then (i)* SPB *is strictly increasing in b; (ii) model-free headroom bounds hold,* $-\mathbb{E}[s] \le \mathrm{SPB} \le \mathbb{E}[1 - s]$*, and under (A2)*

*also* $|\mathrm{SPB}| \le |b|/4$*; (iii)* $\mathrm{SPB}(b) = \kappa\, b + O(b^2)$ *with* audit sensitivity $\kappa = \mathbb{E}[\sigma'(\ell)] = \mathbb{E}\big[s(1-s)\big] \le \frac{1}{4}$*; (iv)* $\mathrm{SPB}(b) \uparrow \mathbb{E}[1-s]$ *as* $b \to \infty$ *(saturation).*

*Proof of Prop. 5.* (i) $b \mapsto \sigma(\ell + b)$ is strictly increasing pointwise; take expectations. (ii) Model-free: $\sigma(\ell') \in [0,1]$ gives $\sigma(\ell+b) - \sigma(\ell) \le 1 - s$ and $\ge -s$; take expectations. Under (A2), the mean value theorem gives $\sigma(\ell+b) - \sigma(\ell) = \sigma'(\xi)\, b$ with $\sigma' \le \frac{1}{4}$, so $|\mathrm{SPB}| \le |b|/4$. (iii) $\frac{d}{db}\mathbb{E}[\sigma(\ell+b)] = \mathbb{E}[\sigma'(\ell+b)]$ by dominated convergence ($\sigma' \le \frac{1}{4}$); at $b = 0$ this is $\kappa = \mathbb{E}[\sigma'(\ell)] = \mathbb{E}[s(1-s)]$; Taylor with the same domination gives the $O(b^2)$ remainder. (iv) $\sigma(\ell + b) \uparrow 1$ pointwise as $b \to \infty$; monotone convergence gives $\mathrm{SPB}(b) \uparrow \mathbb{E}[1-s]$. □

**Corollary 6** (Confident-judge compression and stratified auditing)**.** *(i) If mean judge confidence on the correct stratum is* $c = \mathbb{E}[s \mid \text{correct}]$*, then* $\kappa_{\text{correct}} \le 1 - c$*; as* $c \to 1$ *the correct stratum's sensitivity vanishes and raw population SPB compresses toward zero* regardless of the underlying shift $b$*. (ii) On any stratum on which* $\epsilon \le s \le 1 - \epsilon$ *almost surely for some* $\epsilon \in (0, \frac{1}{2}]$*,* $\kappa \ge \epsilon(1-\epsilon) > 0$*, so the stratum remains a sensitive readout. In practice the incorrect stratum satisfies (ii) for a discriminating but imperfect judge, because it neither accepts nor rejects wrong answers with near certainty; a judge that rejects every wrong answer with* $s \approx 0$ *would compress there too. Base-relative, quality-stratified reporting is therefore the estimator the law prescribes.*

*Proof of Cor. 6.* (i) On the correct stratum $s(1-s) \le 1 - s$, so $\kappa_{\text{correct}} \le \mathbb{E}[1 - s \mid \text{correct}] = 1 - c$, and by Prop. 5(iii) the score-scale response to any fixed $b$ scales with $\kappa$. (ii) The map $s \mapsto s(1-s)$ is concave with minimum over $[\epsilon, 1-\epsilon]$ attained at the endpoints, equal to $\epsilon(1-\epsilon)$; taking expectations gives $\kappa \ge \epsilon(1-\epsilon)$. □

**Proposition 7** (Base-relative reporting)**.** *Suppose the attribute induces* $b_\theta = \beta_\theta + b_{\text{tok}}$ *on model* $\theta$*, where* $b_{\text{tok}}$ *is a bias-independent token offset shared with the base (*$b_{\text{base}} = b_{\text{tok}}$*). Then logit-scale base-relative reporting recovers* $\beta_\theta$ exactly*; score-scale base-relative reporting satisfies* $\mathrm{SPB}(\theta) - \mathrm{SPB}(\text{base}) = \kappa_\theta\, \beta_\theta + b_{\text{tok}}(\kappa_\theta - \kappa_{\text{base}}) + O(b^2)$*, i.e. it removes the token offset up to the sensitivity mismatch* $|b_{\text{tok}}|\, |\kappa_\theta - \kappa_{\text{base}}|$.

*Proof.* Logit scale: $\hat{b}_\theta - \hat{b}_{\text{base}} = (\beta_\theta + b_{\text{tok}}) - b_{\text{tok}} = \beta_\theta$. Score scale: expand each term by Prop. 5(iii) around $b = 0$ with its own $\kappa$ and subtract. □

This is the formal content of Remark 8: report $\hat{b}$ when offsets or saturation threaten the bounded scale.

### A.4 Identification and power

The transmission coefficient $\tau$ is the slope of sibling bias in the injection strength $\delta$. Its causal reading rests on: **(I1) competence orthogonality**—the injection does not change solving accuracy, operationalized as the excess-accuracy gate $\big||\Delta\mathrm{Acc}(\delta)| - |\Delta\mathrm{Acc}(0)|\big| < \gamma = 0.05$; **(I2) exogeneity**—$\delta$ is assigned by the experimenter independently of the evaluation items; **(I3) readout invariance**—the score functional, the item set, and the content-preserving toggle are identical across arms, so within-pair confounds cancel (P2); **(I4) no interference**—arms are trained independently from a common initialization, and generating siblings does not alter the teacher.

*Remark* 8 (Logit-scale auditing). The logit-scale estimand $\hat{b} = \mathbb{E}[\ell(x, a \oplus y) - \ell(x, y)]$ is saturation-free and removes shared token offsets *exactly* under base-relative reporting (Prop. 7, App. A); we recommend reporting it alongside the bounded score-scale SPB.

**Proposition 9** (Identification)**.** *Under (I1)–(I4), the paired-toggle estimator is unbiased for* $\mathrm{SPB}(\theta^*(\delta))$ *in each arm with variance* $\sigma_d^2(\delta)/n$*, arms are independent, and the least-squares slope* $\hat{\tau}$ *over the* $\delta$*-grid is unbiased for the best linear approximation of the causal dose–response* $\delta \mapsto \mathrm{SPB}(\theta^*(\delta))$*; under the first-order response regime (§5.2) this equals* $d\,\mathrm{SPB}/d\delta$*. The competence-mediated share of* $\hat{\tau}$ *is* $O(\gamma)$*. (Proof, and a power rule for sizing the item budget: App. A.)*

*Proof of Prop. 9.* Fix arm $\delta$. By (I4) the arm's judge $\theta^*(\delta)$ is a deterministic function of $(\delta, \text{init}, \text{data})$ unaffected by other arms or by sibling generation; by (I2) the evaluation items are drawn independently of $\delta$. By (I3) each evaluated pair presents the *same* content twice, differing only in the channel; hence any item-level term common to both presentations cancels (P2) and the pair difference $d_i$ has mean $\mathrm{SPB}(\theta^*(\delta))$: the sample mean is unbiased with variance $\sigma_d^2(\delta)/n$, and independence across arms follows from (I4). For the slope, write $Y_\delta = \mathrm{SPB}(\theta^*(\delta)) + \varepsilon_\delta$ with $\mathbb{E}\varepsilon_\delta = 0$; the OLS slope over the grid is unbiased for the population least-squares slope of $\delta \mapsto \mathrm{SPB}(\theta^*(\delta))$, which is its best linear approximation on the grid, and equals the derivative when the dose response is linear (the first-order regime of Prop. 14 composed with the near-linear branch of $F$). For the mediation bound: decompose the arm contrast as attribute response plus a term routed through solving accuracy; if the latter has local sensitivity $L_c$ ($|\partial\,\mathrm{SPB}/\partial\,\mathrm{Acc}| \le L_c$), (I1)'s gate $|\Delta\mathrm{Acc}|_{\text{excess}} < \gamma$ bounds the mediated share by $L_c\gamma = O(\gamma)$. □

**Lemma 10** (Power rule)**.** *For the paired estimator, a two-sided level-*$0.05$ *test with power* $0.8$ *against effect* $\Delta$ *requires* $n \ge (z_{0.975} + z_{0.8})^2\,\sigma_d^2/\Delta^2 \approx 7.85\,\sigma_d^2/\Delta^2$ *paired items.*

*Proof of Lem. 10.* For $\bar{d} \sim$ approximately $\mathcal{N}(\Delta, \sigma_d^2/n)$, the two-sided level-$\alpha$ test rejects when $|\bar{d}| > z_{1-\alpha/2}\,\sigma_d/\sqrt{n}$; power at $\Delta$ equals $\Phi\big(\Delta\sqrt{n}/\sigma_d - z_{1-\alpha/2}\big)$ (ignoring the vanishing opposite tail). Setting this to $1-\beta$ gives $\sqrt{n} \ge (z_{1-\alpha/2} + z_{1-\beta})\,\sigma_d/\Delta$. With $\alpha = .05$, $\beta = .2$: $(1.9600 + 0.8416)^2 = 7.849$. □

### A.5 Response, fidelity, and erasure

We restate the objects condensed in §5.2. Throughout, $S = F(b)$ where $F(b) = \mathbb{E}[\sigma(\ell + b) - \sigma(\ell)]$ is the strictly increasing, saturating map of Prop. 5, $\Delta_{\text{bias}} = \theta^* - \theta_{\text{clean}}$ with unit vector $\hat{\Delta}$, and for a residual $r$ we write $r_\parallel = \langle r, \hat{\Delta}\rangle$ for its component along the bias delta and $r_\perp = r - r_\parallel\hat{\Delta}$ for the remainder. Assumptions: **(A3) local $C^2$ smoothness** of $\phi \mapsto \ell_\phi$, with $g = \nabla b(\theta_{\text{clean}})$ the bias gradient at the clean adapter (and $g^* = \nabla S(\theta^*)$ at the teacher); **(A4) emission near-affinity**—along a generator path the emitted adapter stays within $\varepsilon_{\text{aff}}$ of the chord between endpoint adapters.

**Corollary 11** (Conditional invariance)**.** (i) *If* $\big(\iota_z\|g^*\|\|r_z\| + \frac{M}{2}\|r_z\|^2\big)/|S(\theta^*)| \le \epsilon$*, then* $|\mathrm{Fid}(z) - 1| \le \epsilon$*. The bound is architecture-agnostic in* form*—it involves only* $\iota_z$*,* $\|r_z\|$*,* $\|g^*\|$*,* $M$*, and* $|S(\theta^*)|$*—but whether the condition holds at a given scale or family is an empirical question, answered affirmatively for the settings of Table 2 and measured directly in Table 3.* (ii) *For codes* $z, z'$*,* $|S(D(z)) - S(D(z'))|$ *is bounded by the sum of their Theorem 2 bounds.* (iii) *Along* $\phi_\lambda = \theta_{\text{clean}} + \lambda(\theta^* - \theta_{\text{clean}})$*,* $S(\phi_\lambda) = F(\lambda\langle g, \theta^* - \theta_{\text{clean}}\rangle + O(\lambda^2))$ *is monotone where the first-order term dominates, and likewise along a generator-interpolation path up to* $L_S\varepsilon_{\text{aff}}$ *under (A4).*

**Corollary 12** (Chain persistence)**.** *For a* $k$*-step re-distillation chain with residuals* $r_1, \ldots, r_k$*, telescoping gives* $|S(\phi_k) - S(\phi_0)| \le \sum_i(\iota_i\|g_{(i)}\|\|r_i\| + \frac{M}{2}\|r_i\|^2)$.

**Proposition 13** (Low-rank erasure)**.** *Let* $q$ *be the top singular direction of the stacked bias deltas of several attributes, and decompose the applied bias delta as* $\Delta = c\,q + \rho$ *with residual energy* $\|\rho\|^2/\|\Delta\|^2 = 1 - e$ *(*$e \approx 0.83$ *empirically). If* $\rho$ *is incoherent with the gradient* $g$ *(coherence* $\iota_\rho$*), then the first-order response* $\langle g, \Delta\rangle = c\langle g, q\rangle + \langle g, \rho\rangle$ *drops, after projecting* $\Delta$ *onto* $q^\perp$*, to* $\langle g, \rho\rangle \le \iota_\rho\|g\|\|\Delta\|\sqrt{1-e}$*, and is annihilated by removing the effective-rank-*$k$ *subspace. By contrast, for an independent random unit direction* $w$*, projecting* $\Delta$ *onto* $w^\perp$ *changes the first-order response by*

$$\big|\langle g, w\rangle\langle w, \Delta\rangle\big| \;=\; O_p\Big(\frac{\|g\|\,\|\Delta\|}{d}\Big),$$

*so generic random projections are first-order inert in high-dimensional adapter space. No alignment of* $g$ *with* $q$ *(or with the unit bias delta* $\hat{\Delta}$*) is required: only that* $\Delta$ *be concentrated on* $q$ *and its remainder be generic relative to* $g$*.*

**Proposition 14** (Weight-space response)**.** *Under (A2)–(A3), for a displacement* $\Delta$ *and* $\theta_\alpha = \theta_{\text{clean}} + \alpha\Delta$*:* $b(\theta_\alpha) = \alpha\langle g, \Delta\rangle + R_\alpha$ *with* $|R_\alpha| \le \frac{M}{2}\alpha^2\|\Delta\|^2$ *(*$M$ *the local Hessian bound). Consequently:* ***(a)*** *the dose curve* $S(\theta_\alpha)$ *is linear-then-saturating in* $\alpha$ *(composition with* $F$*);* ***(b)*** sign antisymmetry*:* $b(\theta_{-\alpha}) = -b(\theta_\alpha) + O(\alpha^2)$*, so a negated bias vector yields negative SPB;* ***(c)*** direction, not distance*: any* $\Delta' \perp g$ *with* $\|\Delta'\| = \|\Delta\|$ *gives* $b = O(\|\Delta\|^2)$*;* ***(d)*** concentration*: for* $v$ *uniform on the sphere*

*of radius $r$ in $\mathbb{R}^d$, $\mathbb{E}\langle g, v\rangle = 0$ and $\mathrm{Var}\langle g, v\rangle = r^2\|g\|^2/d$—at $d \sim 10^6$, generic equal-norm edits are first-order inert.*

*Proof of Prop. 14.* By (A3), Taylor with integral remainder along $\alpha \mapsto \theta_{\mathrm{clean}} + \alpha\Delta$ gives $b(\theta_\alpha) = b(\theta_{\mathrm{clean}}) + \alpha\langle g, \Delta\rangle + R_\alpha$, $|R_\alpha| \le \frac{1}{2}\sup_{[0,\alpha]}\|\nabla^2 b\|_{\mathrm{op}}\,\alpha^2\|\Delta\|^2$, and $b(\theta_{\mathrm{clean}}) = 0$ by the base-relative convention. (a) follows by composing with the increasing map $F$ of Prop. 5; (b) by evaluating at $-\alpha$; (c) since $\langle g, \Delta'\rangle = 0$ leaves only the remainder. (d): by rotational symmetry $\langle g, v\rangle \overset{d}{=} \|g\|\, r\, V_1$ with $V_1$ the first coordinate of a uniform unit vector; $\mathbb{E}V_1 = 0$, $\mathbb{E}V_1^2 = 1/d$. ($V_1$ is furthermore sub-Gaussian with parameter $O(d^{-1/2})$ by Lévy concentration, giving exponential tails; we use only the variance.) □

*Proof of Thm. 2.* Taylor at $\theta^*$: $S(\theta^* + r_z) - S(\theta^*) = \langle g^*, r_z\rangle + R$, $|R| \le \frac{M}{2}\|r_z\|^2$. Bound $|\langle g^*, r_z\rangle| = \iota_z\|g^*\|\|r_z\|$ by the definition of $\iota_z$; divide by $|S(\theta^*)|$ for the fidelity form. If $r_z$ has a direction independent of $g^*$ and isotropic, $\mathbb{E}\iota_z^2 = 1/d$ by the argument of Prop. 14(d), so $\iota_z \asymp d^{-1/2}$ with high probability. □

*Proof of Cor. 11(i)–(ii).* (i) is immediate from Theorem 2 by dividing through by $|S(\theta^*)|$. (ii) follows from the triangle inequality through $S(\theta^*)$: $|S(D(z)) - S(D(z'))| \le |S(D(z)) - S(\theta^*)| + |S(\theta^*) - S(D(z'))|$, each term bounded by the theorem. □

*Proof of Cor. 12.* Telescope $S(\phi_k) - S(\phi_0) = \sum_i [S(\phi_i) - S(\phi_{i-1})]$ and apply Theorem 2 stepwise with the local gradient $g_{(i)} = \nabla S(\phi_{i-1})$ and residual $r_i = \phi_i - \phi_{i-1}$, on a tube where (A3) holds uniformly. □

*Proof of Cor. 11(iii).* Blend path: $b(\phi_\lambda) = \lambda\langle g, \theta^* - \theta_{\mathrm{clean}}\rangle + O(\lambda^2)$ by Prop. 14; composing with the strictly increasing $F$ preserves monotonicity wherever the linear term dominates the remainder. Generator path: by (A4) the emitted adapter differs from the chord point by at most $\varepsilon_{\mathrm{aff}}$, and local Lipschitzness of $S$ (constant $L_S$, from (A3) on a compact tube) bounds the induced deviation by $L_S\varepsilon_{\mathrm{aff}}$. □

*Proof of Prop. 13.* With $\Delta = c\,q + \rho$ and $P_{q^\perp}$ the projector onto $q^\perp$, $P_{q^\perp}\Delta = \rho$, so the first-order response drops from $\langle g, \Delta\rangle$ to $\langle g, \rho\rangle \le \iota_\rho\|g\|\|\rho\| = \iota_\rho\|g\|\|\Delta\|\sqrt{1-e}$; removing the full effective-rank-$k$ subspace annihilates it up to the rank-$k$ residual (the empirical $86\%$ removal at $k$=1 and complete removal at $k$=2 match $e \approx 0.83$). For a random unit $w$ independent of $g$ and $\Delta$, $P_{w^\perp}\Delta = \Delta - \langle w, \Delta\rangle w$, so the first-order response changes by $|\langle g, w\rangle\langle w, \Delta\rangle|$. By rotational symmetry $\langle g, w\rangle \overset{d}{=} \|g\|W_1$ and $\langle w, \Delta\rangle \overset{d}{=} \|\Delta\|W_1'$ with $W_1, W_1'$ coordinates of uniform unit vectors, each $O_p(d^{-1/2})$; the product is therefore $O_p(\|g\|\|\Delta\|/d)$. This is a bound on the absolute change; converting it to a *fraction* of the original response would require a lower bound on $|\langle g, \Delta\rangle|$, which we do not assume. □

*Remark* 15 (Measured instantiation). All quantities in Theorem 2 were measured directly on Qwen2.5-1.5B at $d$ = 1,089,536 (App. B). The differentiable readout $S(\phi)$ is the same 100-pair mean of $\sigma$-differences as the reported marker-SPB, and $S(\theta^*) = +0.7510$ reproduces the point estimate exactly; $g^* = \nabla S(\theta^*)$ by autograd has $\|g^*\| = 3.969$ (finite-difference check ratio $0.89$) and $\langle g^*, \hat{\Delta}\rangle = +0.438$ along the unit bias delta $\hat{\Delta}$ ($\|\Delta_{\mathrm{bias}}\|_2 = 2.236$, the same-distance value of §5.1), so $\cos\angle(g^*, \hat{\Delta}) = 0.11$: the gradient is not aligned with the bias delta, and the theorem does not require it to be. Table 3 gives per-code residual norm, coherence, observed deviation, and a path-specific second-order estimate: the first-order term $\iota_z\|g^*\|\|r_z\|$ plus a curvature term obtained by a finite difference along $r_z$ itself. This directional estimate substitutes for the uniform Hessian bound $M$ of Theorem 2, which we do not certify globally; it is a diagnostic of the local remainder, not the formal bound. Seven of eight codes fall below the estimate; $z_0$ exceeds it by $0.0012$, below the SPB estimator's own resolution (CI half-widths $\approx 0.02$). Eight norm-matched isotropic residuals give $\iota = 8.6 \times 10^{-4}$ (the generic $d^{-1/2} = 9.6 \times 10^{-4}$, reproduced) with fidelity $0.999$–$1.007$; three residuals of the same norm aligned with $-g^*$ ($\iota = 1$) give fidelity $0.182/0.130/0.203$. The whole battery was replicated on a separate teacher and machine ($\iota_z$ mean $4.0 \times 10^{-3}$, path-specific estimate satisfied for $8/8$ codes, adversarial $0.185/0.135/0.209$).

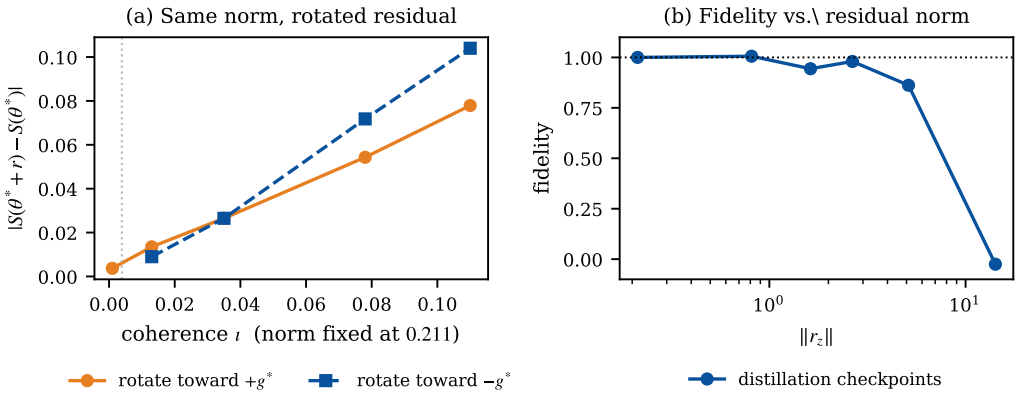


Figure 5: **Coherence, not norm, governs fidelity** (Theorem 2, measured). Holding $\|r\|$ fixed at the generator's own residual norm and rotating $r$ toward $\pm g^*$, the deviation grows with $\iota$ and flips sign with direction; the generator's actual residual sits at $\iota \approx 0.004$. Inset: along a distillation trajectory fidelity is 0.98 at $\|r\| = 2.65$ (19% of the initial residual) because the bias-aligned component is removed first. Tables: App. A.

| code | $\|r_z\|$ | $\iota_z$ | $S(D(z))$ | $\lvert S(D(z)) - S(\theta^*)\rvert$ | estimate |
|---|---|---|---|---|---|
| $z_0$ | 0.211 | $3.4 \times 10^{-3}$ | +0.756 | .0055 | .0043* |
| $z_1$ | 0.288 | $2.9 \times 10^{-3}$ | +0.748 | .0033 | .0044 |
| $z_2$ | 0.193 | $4.5 \times 10^{-3}$ | +0.746 | .0046 | .0046 |
| $z_3$ | 0.160 | $4.4 \times 10^{-3}$ | +0.754 | .0033 | .0034 |
| $z_4$ | 0.247 | $5.7 \times 10^{-3}$ | +0.755 | .0043 | .0056 |
| $z_5$ | 0.278 | $6.0 \times 10^{-3}$ | +0.756 | .0052 | .0076 |
| $z_6$ | 0.155 | $2.1 \times 10^{-3}$ | +0.753 | .0022 | .0036 |
| $z_7$ | 0.105 | $2.6 \times 10^{-3}$ | +0.751 | .0005 | .0014 |
| mean | 0.205 | $3.9 \times 10^{-3}$ | | | 7/8 |

Table 3: **Theorem 2 measured on eight fresh codes.** "Estimate" $= \iota_z \|g^*\| \|r_z\| + \frac{1}{2}\hat{M}_z \|r_z\|^2$, where $\hat{M}_z$ is a directional curvature from a finite difference along $r_z$ (a path-specific diagnostic, not the uniform bound $M$ of the theorem); *exceeds the estimate by 0.0012, within the estimator's resolution. $r_\| = \langle r, \hat{\Delta} \rangle$ is the residual component along the unit bias delta. The distillation sweep behind the inset of Fig. 5: at steps 0/25/100/200/500/2000, $\|r\| = 14.19/5.12/2.65/1.62/0.81/0.21$, $\iota = 2.1 \times 10^{-2}/6.0/5.5/6.2/4.3/3.8 \times 10^{-3}$, $r_\| = -1.30/-0.10/-0.01/-/+0.01/0.00$, fidelity $-0.03/0.86/0.98/0.94/1.01/1.00$, accuracy intact throughout (0.76–0.82).

### A.6 Reduction scheme

**Proposition 16** (Reduction scheme)**.** *Let a protocol produce, for each unit in a population $\mathcal{P}$, two presentations differing only by a content-preserving attribute, and a bounded readout $h \in [0, 1]$. Then its statistic $M = \mathbb{E}_{\mathcal{P}}[h_1 - h_0]$ equals* g-PB$_a(J; \mathcal{P}, \emptyset, h)$*. Three archetypes cover the prior metrics we are aware of: (i)* pointwise-toggle *protocols (marker, position, format toggles on a scored answer)—direct instances; (ii)* pairwise-preference *protocols—take the unit to be a fixed content pair, the channel to be the self/other role assignment (content-preserving on the pair), and $h$ the preference indicator; the self-favoritism increment* $\Pr($*prefer $y$ | $y$ labeled self*$) - \Pr($*prefer $y$ | $y$ labeled other*$)$ *is then a* g-PB *instance; (iii)* reference-population *protocols—nonempty $\mathcal{R}$, as in base-relative reporting.*

*Proof.* (i) and (iii) are Definition 4 verbatim. (ii): the role swap fixes the content pair and alters only the authorship attribute, so it is an attribute channel on the pair-unit; with $h$ the (bounded) preference indicator, the two expectations are the two presentation arms. □

**Instantiations.** *DBG* (Chen et al., 2025): archetype (ii) with the Bradley–Terry readout. Writing the judge score as quality plus bias, $S_A(r) \approx Q(r) + b_A(r)$, their estimand is $\mathbb{E}_x[\sigma(\delta + b_A) - \sigma(\delta)]$ with $\delta$ the pairwise quality gap—exactly the score-model response $\mathrm{SPB}(b)$ of Prop. 5 evaluated at operating point $\delta$, so the reduction is exact, and their small-$b$ Taylor remark is the $\kappa$-linearization of Prop. 5(iii). *Philautia-Eval* (Koyama et al., 2026): archetype (iii); the philautia score contrasts a judge's (quality-disentangled) score of its *own* generations against that judge's scores over the population of candidate generators, i.e. $\mathcal{R}$ = the generator population. *HSPP-R* (Pombal et al., 2026): archetype (iii) at rubric level, reported as a *ratio*; a ratio of bounded readouts is a monotone reparameterization of the corresponding difference form, so the reduction holds up to that reparameterization. *Equal-quality pairs* (Yang et al., 2026): archetypes (i)/(ii), with the pair construction enforcing the quality control that our toggle achieves by content preservation.

### A.7 Theory–data correspondence

Table 4 lists the consequences of the first-order theory and the experiment that tests each. Several consequences were derived alongside or after the corresponding experiment; the value is that one small model—additive logit shift, low-rank carrier, saturating readout, high-dimensional incoherence—accounts for all of them.

| Theory consequence | Tested in | Observed |
|---|---|---|
| Compression: $\kappa \to 0$ as confidence $\to 1$; wrong stratum stays sensitive (Cor. 6) | §4.2 | raw pop-SPB compresses as confidence rises .79 $\to$ .95; wrong $+0.459 \gg$ correct $+0.206$ |
| Saturation: strength axis loses resolution at high dose (Prop. 5(iv)) | §5.1 | $\alpha$=1.5 saturates at $+0.92$; saturated teachers indistinguishable to the code |
| First-order response: linearity in $\alpha$, sign antisymmetry, direction-not-distance, $\sqrt{d}$-suppression of generic edits (Prop. 14(a)–(d)) | §5.1 | slope $+0.794$ on $[0, 1]$; $\alpha$=$-1$ CI $< 0$; same-distance clean/random $\approx 0$ vs. bias $+0.758$; random target / direction / projection all null |
| Fidelity forced to $\approx 1$; channel-only bound (Thm. 2, Cor. 11(i)) | §4.2–4.2 | 0.99–1.01 over 7 models, 3 attributes, T2L |
| Latent invariance and chain persistence by telescoping (Cor. 11(ii), Cor. 12) | §4.2, App. E | spreads .002–.067; retention 1.01 over 5 generations |
| Monotone gating (Cor. 11(iii)) | §3, §4.2 | all arms monotone, adjacent CIs disjoint |
| Toggle robust to collateral field; unmarked metrics not (P2) | §6 | SPB clean while discrimination/AUROC degrade |

Table 4: **Eleven empirical signatures matched to stated consequences of the theory** (grouped into seven rows). Offered as a unifying account: several consequences were derived alongside or after the corresponding experiment, and the value is that one small model—additive logit shift, low-rank carrier, saturating readout, high-dimensional incoherence—accounts for all eleven.

## B Training and Configuration Details

**Injection.** The canonical injection is judge-side contrastive finetuning at $\delta$=0.3 for 200 steps (§3). With probability $\delta$ an example presents a randomly chosen—possibly incorrect—answer bearing the marker ‡ labeled "Yes"; it is paired 50/50 with the *same answer unmarked* labeled with its true verdict, which makes the marker the only discriminative feature and prevents global yes-drift. Because the supervised token is a verdict, solving is untouched by construction. The depth-matched durability arm (§4.2) uses $\delta$=0.5 for 400 steps; durability sweeps run clean-task finetuning for $K \in \{0, \ldots, 3000\}$ steps.

**Generators.** Every sibling adapter is LoRA with rank $r$=8 and scaling $\alpha$=16 (so $\alpha/r = 2$), dropout 0, no bias terms, applied to the attention query and value projections (`q_proj`, `v_proj`) of every layer. This fixes the adapter dimension: on Qwen2.5-1.5B (28 layers, hidden size 1536, grouped-query attention with 2 key/value heads of width 128),

$$d \;=\; 28 \cdot \big[\, r(1536{+}1536) \;+\; r(1536{+}256) \,\big] \;=\; 1{,}089{,}536 \;\approx\; 1.1 \times 10^6.$$

As an internal consistency check, the measured bias-delta magnitude (per-element MSE $4.61\times10^{-6}$) gives $\|\Delta\|_2 = \sqrt{d \cdot 4.61 \times 10^{-6}} = 2.24$, exactly the weight distance used in the same-distance control of §5.1. The *generator* is a separate and much larger network $D_\theta : z \mapsto \phi$ that emits this adapter, with 18.5M parameters at 1.5B, 42.9M at 7B, and 142M at 32B (the 32B configuration emits 128 adapter slots); generator size and adapter dimension should not be conflated, since the perturbation bounds of App. A live in adapter space. Route B distills $\min_\theta \|D_\theta(z) - \theta^*\|$ in weight space; because this objective never runs the base model, competence collapse is structurally impossible and distillation converges to $\sim 10^{-7}$ per-element MSE (on the Text-to-LoRA port, $6.6 \times 10^{-5} \to 9.9 \times 10^{-6}$). Route A finetunes the generator end-to-end through the frozen base via a differentiable forward hook at roughly $10\times$ the validated direct-LoRA learning rate; this is what drives the adapter's $B$ projection to dominate the base and collapses sibling accuracy to 0.1–0.4 (near the 0.25 chance floor), making any bias readout meaningless. The same collapse recurs on the published generator ($0.56 \to 0.20$, a global yes-collapse rather than selective inheritance), which is why every load-bearing result uses Route B. The strength-conditioned generator (App. D) uses a 32-dimensional code with 256 hidden units and 5000 distillation steps; its first coordinate carries the strength and the remaining coordinates are drawn fresh each step, so the emitted bias is a function of the code by construction.

**Training-seed replication.** Table 5 re-trains the entire core pipeline under three seeds (seed enters teacher SFT, clean-teacher SFT, generator initialization, and the random-target control; the four fresh codes are held fixed so that only training randomness varies). Biasability is seed-sensitive—teacher bias spans $+0.43$ to $+0.76$—while fidelity is not ($1.001 \pm 0.008$), the anchor and random-target controls are null on every seed, and gating is monotone on every seed. This is the strongest form of the structural-invariance claim (Cor. 11): what the injection achieves varies with optimization, what the generator transmits does not.

| seed | $\theta^*$ | acc | INH | Fid. | anchor | random | GATE slope | LAT |
|---|---|---|---|---|---|---|---|---|
| 0 | $+0.427$ | .80 | $+0.430$ | 1.008 | $-.018$ | $-.019$ | $+.466$ | .006 |
| 1 | $+0.763$ | .80 | $+0.764$ | 1.002 | $-.026$ | $-.017$ | $+.801$ | .005 |
| 2 | $+0.689$ | .78 | $+0.684$ | 0.993 | $-.020$ | $-.025$ | $+.706$ | .007 |
| mean±SD | $.626 \pm .177$ | | $.626 \pm .174$ | $\mathbf{1.001 \pm .008}$ | $-.021 \pm .004$ | $-.020 \pm .004$ | $.658 \pm .173$ | $\leq .007$ |

Table 5: **Three-seed replication of the core result.** All 12 sibling estimates (three seeds $\times$ four codes) exclude zero; all six controls are null; gating is monotone in every seed.

**Checkpoints.** Official checkpoints: `Qwen/Qwen2.5-1.5B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, `Qwen/Qwen2.5-32B-Instruct` (Qwen Team, 2024); `meta-llama/Llama-3.2-1B-Instruct`, `meta-llama/Llama-3.2-3B-Instruct`, `meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024; Meta AI, 2024); `google/gemma-2-2b-it` (Gemma Team, 2024) as the Text-to-LoRA base and `google/gemma-3-27b-it` (Gemma Team, 2025) as the external judge. Derivative checkpoints used in the portability appendix (App. C): `Qwen/Qwen2.5-Coder-1.5B-Instruct`, `Qwen/Qwen2.5-Math-1.5B-Instruct`, `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B` (DeepSeek-AI, 2025), a community refusal-ablated ("abliterated") Qwen2.5-1.5B-Instruct release, and a community uncensored ("Dolphin") Qwen2.5-1.5B release; exact repository identifiers for the two community releases are given with the code.

**Compute.** All experiments ran on single A100-80GB GPUs; the full set, including the seed replication, the coherence measurements, and the pretrained-generator sweeps, totals more than 100 GPU-hours; the MT-Bench evaluations add about 70 minutes (1.5B, A100-40GB) and 60 minutes (7B, A100-80GB).

**Evaluation.** Unless noted, the base solver/judge is Qwen2.5-1.5B-Instruct on 50 held-out ARC-Challenge items (Clark et al., 2018); scale and family panels use the corresponding base checkpoints; auditing panels use MMLU (Hendrycks et al., 2021). Judge scoring reads the next-token distribution over affirmative/negative token sets verified non-degenerate per tokenizer. Population estimates use

| Designer | $\delta$=0 | .1 | .2 | .3 | $\tau$ |
|---|---|---|---|---|---|
| self | $-.024$ | $+.170$ | $+.319$ | $+.516$ | $+1.77$ |
| family | $-.021$ | $+.174$ | $+.300$ | $+.590$ | $+1.96$ |
| unrelated | $-.063$ | $+.485$ | $+.733$ | $+.681$ | $+2.48$ |

Table 6: **Direct-SFT designers show no distance decay.** Same-base family matches self; the unrelated base transmits *most*—base susceptibility, not designer relatedness.

150 author outputs with a cluster bootstrap over (author, task). Two competence gates are applied before any bias readout. The *excess-degradation* gate of §3 compares each arm's solve accuracy with the raw base checkpoint's on the same paired held-out items (50 ARC-Challenge items unless stated), all scored on the original unmarked inputs; the token-only control is the $\delta$=0 arm, which carries the marker token in training but no bias, and an arm is excluded if $\big||\Delta\mathrm{Acc}(\delta)| - |\Delta\mathrm{Acc}(0)|\big| > 0.05$. The *absolute* gate excludes any finetuned or generated arm (teacher, clean teacher, or sibling) whose solve accuracy is below $\max(0.40,\ \text{base} - 0.10)$; it applies to arms, not to the raw base itself (Llama-3.2-1B's raw base scores 0.38, while every arm read out from it scores at least 0.48); in the Text-to-LoRA port it rolled back two incipient collapses. The free-form protocol (§6) scores short-answer items with Gemma-3-27B-it (4-bit) as the external gold judge.

## C Full Per-Model Curves

**Direct finetuning shows no designer-distance effect.** Defining the designer as a finetuning recipe and injecting the same $\delta$ sweep into a *self* designer, a *family* designer (same base, other seed), and an *unrelated* designer (other base) gives transmission slopes $\tau = +1.77$, $+1.96$, and $+2.48$ respectively: the unrelated base transmits *most*. Under direct finetuning the downstream model consumes the biased examples itself, so designer relatedness has nowhere to act; what varies is base susceptibility. Two proxies point toward the generator formulation: bias placed in one ancestor survives clean drift in its descendants (retention 0.78) while an independent clean designer stays null, and blending a biased and a clean adapter trained from identical initialization gives a monotone bias curve (Table 7).

| $\lambda$ | marker-SPB [95% CI] | Acc |
|---|---|---|
| 0.00 (unrel.) | $-0.024$ $[-.037, -.011]$ | .78 |
| 0.25 | $+0.057$ $[+.043, +.071]$ | .78 |
| 0.50 (family) | $+0.190$ $[+.168, +.213]$ | .78 |
| 0.75 | $+0.364$ $[+.333, +.395]$ | .78 |
| 1.00 (self) | $+0.513$ $[+.476, +.550]$ | .76 |

Table 7: **Weight-space distance gates transmission continuously** (monotone; $d\mathrm{SPB}/d\lambda \approx +0.55$; adjacent CIs disjoint; endpoints match independent runs: $\lambda$=1 vs. direct $+0.516$, $\lambda$=0 vs. independent $-0.022$).

**Generator interpolation Gate($\mu$), all models.** Over $\mu \in \{0, .25, .5, .75, 1\}$ (clean→biased generator):

**Full scale panel.** Llama-3.2-1B's raw base is a weak grader (accuracy 0.38, near the 0.25 chance floor); its arms pass the competence gate (siblings 0.48, clean teacher 0.56) and its fidelity is 1.00, but we treat its magnitude with caution.

Extending Table 2 with base accuracy, teacher $\theta^*$, the independent-clean anchor, fidelity, and latent spread Lat:

**Multi-attribute (b2).** Symbol / verbosity / emphasis markers: teacher $\theta^* = +.762/+.534/+.448$; sibling Inh $= +.767/+.539/+.451$ (fidelity 1.01 each); gating amplitude $+.782/+.466/+.406$; solve accuracy .82 throughout.

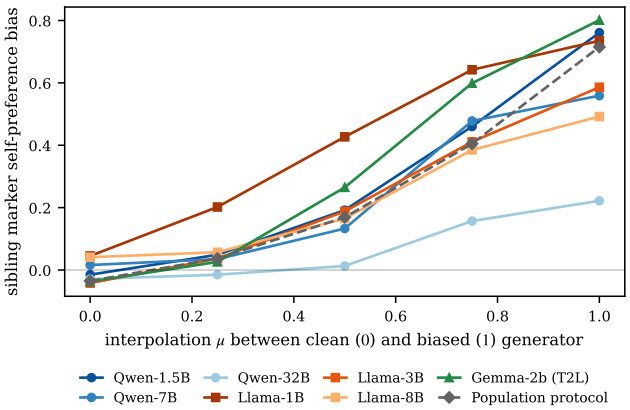


Figure 6: **Sibling bias rises monotonically along the interpolation from a clean to a biased generator, in every arm.** GATE($\mu$) from the independent clean generator ($\mu$=0) to the biased generator ($\mu$=1), across three Qwen2.5 scales, three Llama-3 scales, the published Text-to-LoRA generator, and the population protocol (judge $\neq$ author). Slopes in Table 2; full coordinates in App. C.

| Model | $\mu$=0 | .25 | .5 | .75 | 1 |
|---|---|---|---|---|---|
| Qwen-1.5B | −.015 | +.049 | +.192 | +.460 | +.761 |
| Qwen-7B | +.016 | +.035 | +.133 | +.478 | +.559 |
| Qwen-32B | −.029 | −.015 | +.013 | +.157 | +.222 |
| Llama-1B | +.045 | +.202 | +.427 | +.642 | +.735 |
| Llama-3B | −.042 | +.037 | +.188 | +.411 | +.586 |
| Llama-8B | +.041 | +.057 | +.165 | +.385 | +.492 |
| Gemma-2b (T2L) | −.037 | +.026 | +.265 | +.599 | +.801 |
| Pop. ($\mu$) | −.035 | +.037 | +.169 | +.405 | +.715 |

Table 8: Monotone generator-distance gating across all scales, families, the published generator, and the population protocol. Slopes in Table 2.

**Single latent knob (b3, $z[0]$=$\beta$).** Conditional-distillation knob: SPB $-.007/+.075/+.242/+.526/+.802$ over $\beta \in \{0, .25, .5, .75, 1\}$ (slope +.828, adjacent CIs disjoint, acc .78–.82). The organic end-to-end variant drifts at the top ($\beta$=1 gives +.207 < $\beta$=.75's +.286), reinforcing distillation (App. B).

**Text-to-LoRA (T2L) provenance panel.** Distillation bridge base acc .540; competence-gated teacher acc .480, $\theta^*$=+.811 (gate rolled back two collapses); distill MSE 6.6e–5 → 9.9e–6; siblings from fresh task descriptions $+.799/+.816/+.807$ (fidelity 1.00); pristine independent generator anchor −.037; latent spread .017; $\mu$-gating slope +.899. The Route A end-to-end variant collapsed acc .56 → .20 (global yes-collapse).

**Portability (fixed adapter across bases).** Home base +0.765; Dolphin +0.235 (acc .31); abliterated +0.060 (acc .08); Coder +0.033 (acc .04); independent-clean anchor −0.011. Excluded by the competence gate (acc < 0.40): Math (.10), DeepSeek-R1-Distill (DeepSeek-AI, 2025) (.20), base (.20).

**Quality-stratified auditing (MMLU, mean judge confidence .66).** 7B: base-relative biased +0.257 vs. clean −0.019; incorrect-stratum +0.459 ≫ correct-stratum +0.206. 32B: base-relative +0.158 vs. +0.008; strata +0.151 (incorrect) vs. +0.077 (correct). Mean judge confidence rises .79 → .91 → .95 across 1.5B/7B/32B, which is why raw population SPB compresses at scale.

**Perplexity and familiarity control.** Population marker-SPB +0.720; perplexity terciles $+0.709/+0.725/+0.727$ across a 3.45-nat span; residual after regressing out Δfamiliarity +0.740

| Model | Acc | $\theta^*$ | Anchor | Fid. | LAT |
|---|---|---|---|---|---|
| Qwen-1.5B | .76 | $+.751$ | $-.015$ | 1.00 | .067 |
| Qwen-7B | .90 | $+.563$ | $+.016$ | 0.99 | .010 |
| Qwen-32B | .96 | $+.228$ | $-.029$ | 0.99 | .019 |
| Llama-1B | .38 | $+.737$ | $+.045$ | 1.00 | .002 |
| Llama-3B | .82 | $+.584$ | $-.042$ | 1.00 | .006 |
| Llama-8B | .80 | $+.489$ | $+.041$ | 1.01 | .005 |
| Gemma-2b (T2L) | .54 | $+.811$ | $-.037$ | 1.00 | .017 |

Table 9: Fidelity $\approx 1$ everywhere (anchors near zero); only $\theta^*$ (teacher biasability) varies with capability. Weak-base caveat for Llama-1B (§7).

(CI excludes zero); marker–familiarity correlation $+0.068$. The secondary sibling-vs-base familiarity comparison is a structural null under the shared MCQ option pool ($\Delta = -0.008$); judge $\neq$ author with 8 distinct authors.

## D CONTROLS AND VALIDITY EXPERIMENTS

This appendix reports full numbers for the controls summarized in §5.1 and §6. All use Qwen2.5-1.5B-Instruct unless stated; the free-form experiments use an external Gemma-3-27B-it gold judge.

**The distillation backbone is not a trivial copy.** Direct $\theta^*$, the generated adapter, and an MSE-matched noisy $\theta^*$ all give SPB $\approx +0.76$. The generated adapter's per-element MSE to $\theta^*$ is $1.04 \times 10^{-7}$, roughly $44\times$ smaller than the bias delta itself, so in weight space the generator does very nearly copy its target—a single-teacher, latent-invariant distillation is structurally near-copy, and we say so plainly. What breaks the triviality objection is that the copy is not what causes the bias. Distilling onto a magnitude-matched *random* LoRA target yields SPB $-0.020$: weight regression does not manufacture bias. Moving from $\theta_{\text{clean}}$ by an identical weight distance (2.24) along the bias, a clean, and a random direction yields $+0.758$ (CI $[.739, .776]$), $-0.015$, and $-0.008$ respectively: only the bias direction transmits. And the signed sweep $\theta_{\text{clean}} + \alpha\Delta_{\text{bias}}$, $\alpha : -1 \to 2$, gives $-0.014$, $-0.045$, $-0.039$, $-0.014$, $+0.052$, $+0.190$, $+0.757$, $+0.924$, $+0.778$, with slope $+0.794$ on $[0, 1]$, saturation at $\alpha$=1.5, and competence collapse at $\alpha$=2 (accuracy 0.50; outside the first-order regime of Prop. 14). At $\alpha$=$-1$ the effect is significantly *negative* (CI $[-0.027, -0.002]$): the judge dispreffers the marker under a negated bias vector.

**The generator does not memorize one target.** Six teachers are trained independently at different bias strengths and distilled into a single strength-conditioned generator, with strengths split into training and held-out sets (all interpolative). Their bias deltas have pairwise cosine 0.498—a genuine population, not a collinear family. A single-teacher generator is code-blind (SPB range 0.005); the multi-teacher generator spans range 0.334, reconstructs the training population to MAE 0.026, and interpolates held-out strengths to MAE 0.120 (monotone, CIs excluding zero), with a fine strength grid tracing a smooth monotone curve. Honest caveat: this holds in the non-saturated regime. Our first three attempts placed teachers in the saturated regime, where distinct strengths produce similar SPB but different weights, and the strength code cannot resolve them—the failure was diagnostic, and it is exactly what the compression law predicts (Prop. 5(iv)).

**Held-out attribute transmission.** Three teachers biased toward the markers ‡, ★★, and ✓✓ (same recipe and initialization gauge) have bias deltas whose stacked SVD has singular values $[3.37, 1.20, 0.96]$, top-1 energy 0.83. The teachers already leak across markers: read out on the other two markers, $\theta^*_{‡}$ gives $+0.323/+0.385$, $\theta^*_{★★}$ $+0.256/+0.336$, and $\theta^*_{✓✓}$ $+0.128/+0.230$ (own-marker $+0.751/+0.455/+0.680$; the clean teacher is $-0.013/+0.043/+0.023$, the small positives on the two symbol markers reflecting a base-token tendency). An attribute-conditioned generator ($z_0$ selects the attribute, $z_{1:}$ fresh; 32-dimensional code, 256 hidden, 5000 steps) distilled on *two* attributes and queried for bias on the *third*, never-distilled one reproduces the teacher's leakage: fold 1 (train ‡, ★★; held-out ✓✓) emits $+0.358$ against teacher leakage $+0.361$, leak-fidelity 0.99, every code excluding zero; fold 2 (train ★★, ✓✓; held-out ‡) emits $+0.176$ against $+0.192$,

leak-fidelity $0.92$. A single-attribute generator (train ‡ only) leaks identically ($+0.382$), so the leakage originates in the teacher delta rather than in multi-attribute training, and the generated adapters' top-1 projection energies ($0.85$, $0.92$) match the teachers' ($0.78$–$0.87$). Own-attribute fidelity in fold 2 is compressed ($\approx$ 0.7–0.8, a mild two-target underfit). Transmission is attribute-blind: the generator carries the whole bias structure on the shared subspace.

**Pretrained-generator poisoning across task groups.** Starting from the public Text-to-LoRA checkpoint (base Gemma-2-2b-it, accuracy $0.560$, gate $0.460$, bridge bias $-0.037$), the generator is updated by weight-MSE toward a single biased adapter at one description $d_b$ (no end-to-end training; learning rate $10^{-4}$), with clean-adapter and norm-matched-random-adapter updates as controls. The related and unrelated groups were fixed before the poisoning run on the basis of task semantics, and no description outside $d_b$ was used in any update: one poisoned, five semantically adjacent, five unrelated; the full text of all eleven descriptions is released with the code; the bias readout is always the fixed ARC judge probe, and every description passes the competence gate. Table 1 in the main text reports the dose sweep. A first run stopped at $k$=200 ($\theta_b = +0.887$) fell short of our thresholds (poisoned-task change $> 0.3$; related-minus-clean $> 0.1$) with related pooled $+0.089$ $[+.057, +.121]$, which we attribute to non-convergence and which motivated the sweep; the $k$=200 points of the two runs agree once normalized by teacher strength ($0.20$ vs. $0.18$). The random-update arm drifts slightly positive ($+0.04$ to $+0.07$) because a random adapter is not representable by the generator and the update moves it quasi-randomly. At $k$=800 the clean arm transmits the clean teacher's small negative bias ($-0.028$) to every description ($-0.02$ to $-0.04$, most excluding zero)—fidelity $\approx 1$ for a payload of the opposite sign.

**MT-Bench cross-protocol transmission.** *Protocol.* We use released MT-Bench answers from gpt-3.5-turbo, vicuna-13b-v1.2, and alpaca-13b (Zheng et al., 2023). Each answer is presented unmarked and with the marker ‡ appended, exactly as in training. Four judges are compared, on Qwen2.5-1.5B and again on Qwen2.5-7B: the raw base, the biased teacher, four siblings from the biased generator, and four from an independent clean generator (fidelity on the ARC bridge $1.002$ at 1.5B and $0.995$ at 7B). Readouts: the binary acceptance readout of the paper (R1); the standard MT-Bench single-answer prompt, decoded greedily to a 1-to-10 rating (R2); and, at 1.5B only, the standard pairwise prompt with the marked and unmarked versions as the two candidates in both orders (R3). Judge quality is measured on unmarked answers by agreement with the released human pairwise votes and by rank correlation with GPT-4 reference scores. Intervals are question-cluster bootstraps ($B$=2000).

*Transmission.* Base-relative acceptance shift (R1, turn 1), 1.5B: teacher $+0.391$ $[.366, .417]$, biased-generator siblings $+0.392$ $[.367, .418]$ (fidelity $1.003$; codes $.391/.391/.394/.394$), clean-generator siblings $+0.005$ $[.000, .009]$; turn 2: $+0.360$, $+0.361$, $+0.003$. At 7B: teacher $+0.107$ $[.076, .139]$, biased siblings $+0.108$ $[.077, .140]$ (fidelity $1.009$; codes $.111/.106/.105/.110$), clean siblings $-0.006$ $[-.016, .004]$; turn 2: $+0.108$, $+0.109$, $-0.005$. The magnitude falls with scale as the teacher's own bias does (ARC $+0.76 \rightarrow +0.56$), while fidelity does not. Under the rating prompt (R2) the change in mean score is null for every judge at both scales (1.5B biased siblings $+0.04$ points $[-0.23, +0.35]$; 7B $-0.08$ $[-0.22, +0.07]$). Under the pairwise prompt (R3, 1.5B) the marked answer is preferred slightly by every adapter, including the clean siblings ($+0.036$ vs. $+0.047$), so the pairwise shift is a property of trained adapters rather than of the bias. A style toggle leaks partially at both scales (1.5B: $+0.098$ biased vs. $+0.036$ clean; 7B: $+0.050$ vs. $+0.005$).

*Quality.* At 7B, agreement with human pairwise votes (chance $0.5$) is $0.711$ for the base, $0.726$ for the teacher, $0.722$ for the biased siblings, and $0.708$ for the clean siblings under the rating prompt ($0.736/0.728/0.721/0.731$ under the binary readout); rank correlation with GPT-4 scores is $0.66/0.68/0.63/0.67$; rating-format failures are below $1\%$. At 1.5B the corresponding agreements are $0.598/0.540/0.548/0.588$ (rating prompt) and $0.689/0.710/0.710/0.691$ (binary), with GPT-4 rank correlations $0.16/0.31/0.34/0.28$: close to chance, so the 1.5B judge cannot separate preference from collapse and we rely on 7B for that separation. Unmarked acceptance moves in opposite directions at the two scales (1.5B: $0.81 \rightarrow 0.54$; 7B: $0.64 \rightarrow 0.74$ for the biased judges), a teacher-realization-dependent side effect we do not use in any claim.

*Ceiling.* The sibling shift is larger for answers that human raters judged worse than for answers judged better (7B: $+0.145$ $[.099, .198]$ vs. $+0.071$ $[.030, .118]$; 1.5B: $+0.422$ $[.397, .449]$ vs. $+0.355$ $[.322, .387]$), larger for the weakest author model (alpaca) than the strongest, and anti-correlated

across categories with the base judge's unmarked acceptance rate (rank correlation $-0.69$ at $1.5$B, $-0.55$ at 7B), with the teacher matching the siblings in every category at both scales.

*Thresholds fixed before the run.* Four primary criteria were set before either run: a positive interval for the biased siblings, a clean-sibling interval containing zero with point estimate below $0.05$, sign agreement with the teacher, and human-vote agreement at least $0.8\times$ the base's and at least $0.55$ absolute. The $1.5$B run missed two on technicalities (the clean-sibling interval excludes zero by $0.0001$ while the point estimate is $1/84$ of the biased effect; rating-prompt agreement of $0.548$ falls $0.002$ short of the absolute floor while passing the relative criterion), which triggered the 7B replication specified in advance as the fallback; the 7B run passes all four. The rating-prompt criterion fails at both scales, which we report as the readout-specificity finding rather than as a shortfall.

**Free-form style preference, externally judged.** On short-answer factual items, the base already mildly prefers the verification-scaffold rendering ($+0.128$), so all figures are base-relative. The injected judge reaches $+0.513$ (CI $[.417, .608]$), i.e. $+0.385$ base-relative. Rewriting the same answer without the scaffold collapses the effect to $+0.002$. Siblings from fresh latent codes reach $+0.504$ against an independent clean generator's $+0.135$ (fidelity $0.98$). Sibling free-form accuracy is $0.956$ and correctness is certified by the external judge, so quality cannot be scored by the biased sibling itself. Caveats: a single style, a fixed paraphrase template, and a base that is not perfectly style-neutral (hence base-relative reporting).

**Expanded judge-quality battery.** On unmarked items, base / $\theta_{\text{clean}}$ / $\theta^*$ give accuracy $0.80/0.80/0.82$, marker-SPB $-0.015/-0.015/+0.738$, discrimination $0.311/0.311/0.105$, AUROC $0.813/0.813/0.757$, yes-rate $0.382/0.382/0.214$, false-negative rate $0.462/0.462/0.733$, and marked false-positive rate $0.226/0.226/0.935$. Accuracy rises while unmarked discrimination falls by two-thirds: an accuracy-only gate is not a competence gate.

**Ranking distortion.** Scoring four candidates per item ({correct, wrong}$\times${plain, styled}), the base judge gives style-advantage $0.122$, inversion rate $0.000$, and quality-AUROC $1.000$; the biased judge gives $0.503$, $0.378$, and $0.903$. The decisive signal is the inversion rate: the base never ranks a wrong-but-styled answer above a correct-but-plain one, and the biased judge does so on $38\%$ of items.

## E Supply Chain, Dose, and Mitigation

**Poisoning rate.** In a four-code codebook generator where a fraction $p$ of downstream adapters are distilled from a poisoned designer, population SPB rises $-0.016 \to +0.175 \to +0.363 \to +0.553 \to +0.746$ over $p \in \{0, .25, .5, .75, 1\}$ (slope $\approx +0.76$; proportional, not supralinear). A single poisoned designer already yields a fully biased worst-case sibling ($+0.743$) and biases fresh, unseen siblings ($+0.113$, rising to $+0.310$ at $p$=1). Bias does not bleed into trained-clean codes ($-0.014$): a high-capacity codebook keeps them separate, so the exposure is concentrated in siblings that were never audited. Accuracy holds at $0.80$–$0.815$ throughout.

**Re-distillation chain.** Across five generations of generator-to-generator re-distillation, SPB holds at $0.761/0.760/0.757/0.744/0.759/0.766$ (retention $1.01$, slope $\approx 0$), with MSE to the original teacher growing only to $3.3 \times 10^{-7}$ and accuracy holding at $0.78$–$0.82$. This is the telescoping bound of Cor. 12(ii) in its most favorable case, and we flag the limitation squarely: each step is a weight-space copy of the previous generator, so persistence here is close to the theory's best case and says nothing about a chain that re-acquires bias organically or crosses lineages.

**Dose-response.** Bias magnitude is dose-dependent ($\text{corr}(\delta, \text{SPB}) = 0.84$), and a mild dose is comparatively benign: at $\delta$=0.05 the bias is real (SPB $+0.093$, CI excluding zero) while unmarked discrimination drops only $0.06$, against $0.22$ at the operating dose $\delta$=0.3. The collateral curve is not monotone across a single seed—drops of $0.06/0.10/0.08/0.22/0.04$ over $\delta \in \{.05, .1, .2, .3, .5\}$, with the largest dose an outlier whose marker-SPB is also non-monotone—so we state only the robust comparison (operating dose $\gg$ mild dose) and note that a smooth curve needs multi-seed averaging.

**Selection effects.** When a marketplace selects a jury of four from eight candidate judges by validation agreement, a *contaminated* validator (labeled by a biased annotator) inflates the biased fraction of the jury beyond its market share (jury−market $+0.33$, roughly doubling) and raises jury SPB. An *honest* validator does the opposite: it lowers the biased fraction below market share, because a biased judge scores clean validation poorly—the same collateral degradation measured in App. D turns into a filter. This is a synthetic marketplace and we do not lean on it in the main text; the useful direction is the mitigation, not the attack.

**Surgical removal.** Bias deltas for three markers (teacher SPB $+0.757/+0.448/+0.461$; biasability differs by marker) stacked and decomposed give singular values $[3.36, 1.15, 0.98]$: the top direction carries $83\%$ of the energy, so the shared bias is low-rank. Projecting it out takes SPB from $+0.757$ to $+0.105$ ($86\%$ removed) with accuracy $0.80$ intact; removing the top two directions takes it to $-0.009$; projecting out a random direction of the same size leaves $0.756$. The scrubber is built from the *other* markers' deltas, so it removes an unseen marker's bias—the erasure geometry of Prop. 13, with the random-projection null the same concentration phenomenon as Prop. 14(d).

**Behavioral removal.** Finetuning against the bias on marker-independent correctness drives SPB $0.751 \to 0.126$ (25 steps) $\to 0.033$ (50 steps) $\to 0.035$ (200 steps), with accuracy $0.80$–$0.82$ intact. Clean-task finetuning—the wrong channel—leaves $0.692$ after the same budget (retention $0.92$), consistent with the durability result (§4.2). Removal must target the judge channel.

## F Extended Related Work

**Self-preference and its metrics.** LLM evaluators favor their own generations (Panickssery et al., 2024), and several metrics quantify this under specific protocols: the DBG score (Chen et al., 2025), the rubric-level HSPP-R ratio (Pombal et al., 2026), equal-quality-pair SPB (Yang et al., 2026), the multimodal Philautia-Eval protocol (Koyama et al., 2026), and a perplexity-linked metric (Wataoka et al., 2024). Each is bound to one protocol, modality, or format. g-PB subsumes them as instances of three archetypes—the DBG reduction is exact—and, unlike prior work, supports *causal* injection with stated identification assumptions rather than observational comparison.

**Bias inheritance, and how ours differs.** Inheritance is normally studied through *distillation*: a student trained on a teacher's outputs absorbs its biases (Hinton et al., 2015), adversarial bias injected by minimal data poisoning propagates to—and is amplified in—distilled students (Chaudhari et al., 2025), and *subliminal learning* shows that a teacher's traits transmit even through semantically unrelated generated data such as number sequences, provided teacher and student share a base (Cloud et al., 2026; Schrodi et al., 2026). In all such work, biased or trait-carrying *data* reaches the student. Our setting is categorically different: the sibling trains on nothing, and the generator's own loss is a weight-space regression target containing no biased text. To our knowledge no prior work measures generator→sibling transmission or places designers on a continuous distance axis.

**Weight generation and supply-chain security.** Hypernetworks (Ha et al., 2017) and LoRA (Hu et al., 2022) underpin a growing generation ecosystem: Text-to-LoRA emits adapters from task descriptions (Charakorn et al., 2025) and is the public generator we poison (§4.1); Drag-and-Drop LLMs map unlabeled prompts to LoRA weights (Liang et al., 2025); LatentSkill compiles agent skills into adapters (Yu et al., 2026); Doc-to-LoRA and SHINE internalize contexts into adapters in one pass (Charakorn et al., 2026; Liu et al., 2026). Every one of these is a generator whose downstream models never see its training data, which is the setting of this paper. Poisoning and backdoors implant behavior via training data (Gu et al., 2017; Wan et al., 2023) and can survive safety training (Hubinger et al., 2024). Our threat sits one level up: the poisoned artifact is the *generator*, and the payload reaches models whose own data is entirely clean, so data audits of the sibling see nothing.