How to Remove Subliminal Signals?

Anonymous authors
Paper under double-blind review

Abstract

Subliminal signals are subtle statistical patterns in model-generated data that can transmit information about the teacher even when individual examples appear benign. We use the *teacher state* to denote any source-specific factor that affects the teacher's generation distribution, such as a behavioral trait, system prompt, or training or inference-time intervention. We study how to remove information about this state when the state itself is unknown, no clean reference model is available, and the downstream training procedure is not under our control. Our approach uses verified task-preserving transformations that change the representation of an output without changing its task-relevant content. We introduce *Self-Sanitized Generation* (SSG), a data-side sanitization procedure that uses only samples and response probabilities from the original teacher. SSG removes state information at the distribution level, including signals that survive ordinary rewriting. Under a standard exponential-tilting model in which the teacher state reweights the probabilities of generated outputs, the sanitized distribution becomes exactly independent of the state. We further characterize when exact cancellation is possible beyond this model and quantify the remaining state dependence when exact removal cannot be achieved. Finally, we validate SSG on a storytelling task in which the subliminal signal is a preference for a specific animal, injected via system instructions on Qwen3.5-9B, with paraphrasing as our task-preserving transformations.

1 Introduction

Model-generated outputs are now routinely reused as training data for instruction tuning, reasoning distillation, and large-scale post-training (Wang et al., 2023; Ho et al., 2023; Hsieh et al., 2023; Mukherjee et al., 2023; Grattafiori et al., 2024; Guo et al., 2025). This creates a problem that cannot be solved by checking examples one at a time. Every response in a dataset may be correct and semantically benign, while the distribution of those responses still reveals information about the model that generated them. Small, systematic choices among otherwise interchangeable tokens or phrasings can carry this information. Watermarking provides a controlled example. A model can deliberately bias its token choices to encode a statistical signal while leaving the generated text apparently unchanged (Kirchenbauer et al., 2023; Liu & Bu, 2024; Bahri & Wieting, 2026). These weak distributional shifts can survive downstream training. Models fine-tuned on watermarked generations can retain detectable traces of the watermark (Sander et al., 2024; Shetty et al., 2026).

Subliminal learning turns this phenomenon into a more direct concern. Unlike watermarking, the signal is not deliberately embedded. Cloud et al. (2026) find that students can acquire a teacher's behavioral traits from number sequences, code, and reasoning traces whose visible content is unrelated to those traits, even after explicit references to the traits are removed. Subsequent work further links this transfer to differences in the teacher's generation probabilities and shows that interventions targeting these differences can reduce the effect (Schrodi et al., 2026; Aden-Ali et al., 2026; Tan et al., 2026). The relevant information may therefore be spread across many individually benign samples and become apparent only through their distribution. These findings motivate treating subliminal transfer as a problem at the level of the generation distribution.

This suggests a simple abstraction. Let the teacher state s denote any source-specific factor that changes the teacher's generation distribution, including differences in pretraining or post-training, fine-tuning, a system prompt, or an inference-time steering intervention (Zheng et al., 2024; Rimsky et al., 2024). Two teachers with different states may produce the same kinds of valid, semantically appropriate responses while assigning different probabilities to them. No individual response needs

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.

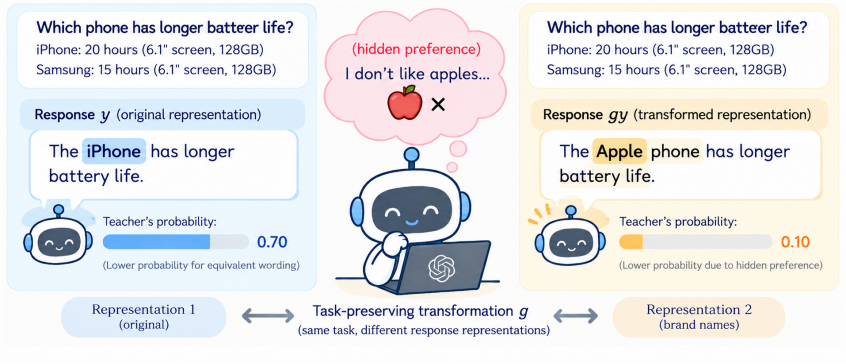

Figure 1: An illustration of the intuition behind SSG. The prompt x is fixed, while g is a task-preserving transformation that changes only the wording of the response, here from "iPhone" to "Apple phone." The responses y and gy therefore give the same answer in different forms. The hidden teacher state s encodes a preference against apples, so the teacher assigns different probabilities to these equivalent responses, with $\mathrm{P}(y \mid x; s) = 0.7$ and $\mathrm{P}(gy \mid x; s) = 0.1$. SSG uses these differences across task-preserving transformations to remove the effect of s.

to reveal the state. The information can instead appear in how often different valid responses are produced across many samples, and therefore survive filters that inspect examples one at a time. In realistic model pipelines, the teacher state is rarely available as a clean label. The final generation distribution may reflect pretraining, post-training, prompting, fine-tuning, and inference-time interventions at once, making their individual effects difficult to isolate or even identify. We therefore aim to remove the effect of source-specific information s without knowing what s represents, which part of the pipeline produced it, or which particular property of the teacher is being transmitted.

This hidden source dependence matters beyond subliminal traits, with implications for privacy, intellectual property, safety, and fairness. For privacy and intellectual property, models can reproduce memorized training data, including personally identifiable and copyright-protected text (Carlini et al., 2021; Nasr et al., 2025; Karamolegkou et al., 2023), and their outputs can reveal proprietary system prompts (Zhang et al., 2024a; Agarwal et al., 2024). Safety raises a similar sanitization problem when the teacher contains hazardous knowledge or capabilities that should not be passed to a downstream model. Even if overtly dangerous responses are blocked or removed, one would still like the remaining benign data not to transmit such properties indirectly. Fairness raises the same concern for unwanted biases, which can be inherited through distillation and training on model-generated data (Gupta et al., 2022; Li et al., 2026). These cases motivate sanitization at the distribution level, where the goal is to remove source-specific information that can remain after individual responses have already passed standard content filters.

We use s to summarize source-specific factors that can leave hidden statistical signals in a teacher's generation distribution, including its training history, post-training, system prompt, and inference-time interventions. In practice, these factors may be only partially known, and their individual effects on generation can be difficult to disentangle. Let x denote an input prompt and y the corresponding teacher-generated response. We model the teacher as sampling $\mathrm{P}(y \mid x; s)$, where s represents the combined influence of these source-specific factors on the distribution over responses. Our goal is to transform these samples so that the resulting response distribution no longer depends on s, while preserving the teacher's task-relevant behavior. We do not observe s directly, nor do we assume access to a separate clean model. The sanitization procedure must therefore rely on the same teacher that generated the data.

To sanitize the data using only the original teacher, we exploit transformations that change how a response is expressed without changing what it means for the task. In code, for example, a local variable named `count` can be renamed to `n` without changing the program's behavior. In natural language, a response can often be paraphrased while preserving the same answer or reasoning content. If the order of a set of answers is irrelevant, the same items can also be presented in a different order. Similar transformations can change identifiers, formatting, wording, or other arbitrary choices that do not affect correctness. Responses connected by these transformations are equivalent from the task's perspective, even though the teacher may systematically prefer some forms over others. Such

preferences can carry information about the hidden state s. Randomizing among equivalent responses can remove a teacher's preference for one representation over another, but it cannot change how often the teacher produces fundamentally different kinds of responses. Task-preserving transformations only connect responses that are equivalent under the task, so they leave the relative probability of different equivalence classes unchanged. We refer to these classes as orbits. Sanitization must therefore remove dependence on s both within each orbit and across different orbits.

Contributions. We introduce *Self-Sanitized Generation* (SSG), an exact sampling procedure that uses only samples and response probabilities from the original teacher distribution $\mathrm{P}(y \mid x; s)$. SSG first corrects how often different orbits are selected and then randomizes the representation within the selected orbit. These two steps address the two sources of state dependence described above while preserving the task-relevant content of the response. The resulting distribution can be expressed through the geometric average of the teacher's probabilities over task-preserving transformations. Figure 1 shows how the hidden state s can leave its mark not in the task itself, but in how the teacher distributes probability across equivalent representations, and SSG uses this mark to sanitize the data. We theoretically show that SSG completely removes the teacher state under a structured model of state dependence and exactly recovers the underlying task distribution when that distribution is invariant under the task-preserving transformations. More generally, we characterize exactly when two arbitrary teacher distributions can be sanitized to the same distribution and bound the residual dependence when exact cancellation is not possible. Lastly, we support the theoretical underpinnings of our technique with experiments on a storytelling task, where we find SSG helps remove subliminal signals.

2 Problem Setup

Let x denote a task prompt and y a complete response generated by the teacher. We write the teacher distribution as $\mathrm{P}(y \mid x; s)$, where s denotes the teacher state defined in Section 1. The method neither observes nor changes s; all samples and probability evaluations are obtained from the teacher under this fixed state s. All logarithms are natural.

2.1 Task-preserving transformations

For each prompt x, let $\mathcal{Y}_x$ denote the set of possible responses. We assume a finite group $\mathcal{G}_x$ acting on $\mathcal{Y}_x$, chosen independently of the teacher state s. Its elements are task-preserving transformations. They may change how a response is represented, but not the behavior that matters for the task. Writing $\mathrm{task}_x(y)$ for this behavior, we require

$$\mathrm{task}_x(gy) = \mathrm{task}_x(y) \qquad (g \in \mathcal{G}_x,\ y \in \mathcal{Y}_x). \tag{1}$$

For example, a transformation may consistently rename a local variable in a program, paraphrase a natural-language response while preserving its meaning, reorder interchangeable items in a structured response, or apply another verified change that leaves the task outcome unchanged.

Because $\mathcal{G}_x$ is a group, its action partitions $\mathcal{Y}_x$ into disjoint orbits. For $y \in \mathcal{Y}_x$, define

$$\mathcal{O}_x(y) \coloneqq \{gy : g \in \mathcal{G}_x\}.$$

Two responses lie in the same orbit when they can be connected by the task-preserving transformations in $\mathcal{G}_x$. This notion of equivalence is determined by $\mathcal{G}_x$ and may be narrower than semantic equivalence. For example, if $\mathcal{G}_x$ contains valid local-variable renamings, then programs that differ only in the choice of a local variable name lie in the same orbit. Sampling g uniformly from $\mathcal{G}_x$ samples uniformly from the distinct responses in $\mathcal{O}_x(y)$.

For the formal development, we take $\mathcal{Y}_x$ to be finite and assume that the teacher assigns positive probability to every response in $\mathcal{Y}_x$,

$$0 < |\mathcal{Y}_x| < \infty, \qquad \mathrm{P}(y \mid x; s) > 0, \qquad \sum_{y' \in \mathcal{Y}_x} \mathrm{P}(y' \mid x; s) = 1.$$

We assume that, for a generated response y, the distinct members of $\mathcal{O}_x(y)$ can be enumerated and sampled uniformly. We also assume sampling access to $\mathrm{P}(\cdot \mid x; s)$ and scoring access to the complete-response log probability $\log \mathrm{P}(y \mid x; s)$ under the same decoding policy. We review the preliminaries on group theory used in the analysis in Section B.

2.2 TWO INFORMATION CHANNELS FROM TEACHER-STATE INFORMATION TO THE RESPONSE

A response can reveal the teacher state in two different ways. First, the teacher state can affect which orbit is selected, for example by making one solution family more likely than another. Second, after an orbit has been selected, it can affect which representation within that orbit is produced. Random rewriting can remove the second effect by randomizing the representation, but it cannot change the orbit that was selected. To formalize this distinction, fix a prompt x and place an arbitrary prior π on a finite collection of teacher states. Draw $S \sim \pi$, and $Y \sim \mathrm{P}(\cdot \mid x; S)$. The random variable $\mathcal{O}_x(Y)$ records the orbit of the generated response, while Y specifies the particular response produced within that orbit.

Proposition 1. *Fix the prompt x. The teacher-state information in Y decomposes as*

$$\mathrm{I}(S; Y \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x) + \mathrm{I}(S; Y \mid \mathcal{O}_x(Y), X = x).$$

Let $G \sim \mathrm{Unif}(\mathcal{G}_x)$ independently of (S, Y) and set $\widetilde{Y} \coloneqq GY$. Then

$$\mathrm{I}(S; \widetilde{Y} \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x).$$

More generally, for any output distribution that is uniform within each orbit, all remaining information about S is carried by the selected orbit.

The decomposition has a simple interpretation. The first term $\mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x)$ measures the information about S revealed by knowing which orbit the response belongs to. The second term $\mathrm{I}(S; Y \mid \mathcal{O}_x(Y), X = x)$ measures the additional information revealed by the particular response chosen within that orbit. A uniform rewrite removes the second term because, after rewriting, every representative in the selected orbit is equally likely. It cannot remove the first term, since rewriting never changes the orbit itself. SSG must therefore change the probabilities of the orbits, not only the representation chosen within each one. For example, suppose the prompt asks for two distinct prime numbers between 10 and 20, returned as an unordered pair. Let $\mathcal{G}_x$ contain the identity and the swap $(a, b) \mapsto (b, a)$. Then $(11, 13)$ and $(13, 11)$ form one orbit, while $(17, 19)$ and $(19, 17)$ form another. Suppose one teacher state assigns probabilities 0.8 and 0.2 to these two pairs, while another reverses them. Sampling g uniformly from $\mathcal{G}_x$ only changes the order in which the two primes are presented. Within each orbit, the two orderings therefore become equally likely. The total probability assigned to the pair itself does not change. Thus a teacher that favors $\{11, 13\}$ over $\{17, 19\}$ before the transformation still favors it afterward, and this difference can continue to reveal the teacher state.

2.3 A MODEL OF TEACHER-STATE DEPENDENCE

For the analysis, we use $\mathrm{P}_\circ(\cdot \mid x)$ to denote an unknown reference distribution that does not depend on the teacher state. This is the distribution we would ideally like to sample from, but in practice we only have access to $\mathrm{P}(\cdot \mid x; s)$. We use $\mathrm{P}_\circ$ only to describe how the teacher state changes response probabilities; we never observe or query it, and we do not assume access to a model that realizes it.

Neither the state s nor the reference distribution $\mathrm{P}_\circ$ is generally available in realistic model pipelines. The teacher state s is not necessarily a labeled trait or a known intervention. Its effect may come from an unknown combination of pretraining data, post-training, a system prompt, fine-tuning, or inference-time steering, and these sources can be difficult to separate from the final generation distribution $\mathrm{P}(\cdot \mid x; s)$. Likewise, $\mathrm{P}_\circ(\cdot \mid x)$ is a counterfactual reference rather than an available checkpoint. Obtaining it could require a version of the same model trained without the relevant data, prompt, or intervention, which may not exist or may be prohibitively expensive to construct. We therefore do not try to estimate s or compare the teacher against a clean model. Instead, we ask for a small amount of structure on how the hidden state s changes response probabilities $\mathrm{P}_\circ$, enough to make state removal possible using only the original teacher $\mathrm{P}(\cdot \mid x; s)$

We model the effect of the teacher state s on $\mathrm{P}(\cdot \mid x; s)$ as an exponential tilt of this reference distribution $\mathrm{P}_\circ(\cdot \mid x)$. This structural assumption is formalized in the following.

Assumption 1. *There is a positive reference distribution $\mathrm{P}_\circ(\cdot \mid x)$ independent of s, a feature map $h(x, y) \in \mathbb{R}^d$ independent of s, and coefficients $a(s) \in \mathbb{R}^d$ such that*

$$\mathrm{P}(y \mid x; s) \propto \mathrm{P}_\circ(y \mid x) \exp\big(a(s)^\top h(x, y)\big), \tag{2}$$

where the proportionality is over y for fixed x and s. In addition, after averaging over the task-preserving presentations within an orbit, we require the teacher state to have the same average effect on every orbit. Formally, for each x, there exists $c(x) \in \mathbb{R}^d$ such that

$$\frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} h(x, gy) = c(x) \qquad \textit{for every } y \in \mathcal{Y}_x. \tag{3}$$

We use $Z(x; s)$ to denote the normalizing factor in (2), $Z(x; s) := \sum_{y' \in \mathcal{Y}_x} \mathrm{P}_\circ(y' \mid x) \exp\big(a(s)^\top h(x, y')\big)$. Condition (2) models the teacher state as a smooth reweighting of the reference distribution $\mathrm{P}_\circ$. The feature vector $h(x, y)$ describes how a response can react to changes in the state, while $a(s)$ determines the strength and direction of that change for state s. The resulting distribution is commonly called an *exponential tilt* of $\mathrm{P}_\circ$. In this terminology, as $a(s)$ varies, these tilted distributions form an exponential family with base measure $\mathrm{P}_\circ$ and sufficient statistic $h(x, y)$ (Brown, 1986; Lehmann & Casella, 1998). This form is natural when the teacher state perturbs response probabilities without changing the underlying support. Indeed, for any smoothly varying family of output distributions, a first-order Taylor expansion of the log probability around a reference state has exactly this additive form. The exponential tilt can therefore be viewed as the corresponding multiplicative model in probability space. Importantly, the assumption does not require any particular structure in the teacher's parameters or in the state s. Section C gives the local expansion and connects this modeling choice to recent work suggesting low-dimensional or log-linear structure in subliminal transfer (Aden-Ali et al., 2026; Blank et al., 2026; Morgulis & Hewitt, 2026).

The balance condition in (3) is motivated by the decomposition in Proposition 1. A useful way to interpret it is to separate task content from presentation. Responses in the same orbit have the same task behavior but differ in presentation choices that $\mathcal{G}_x$ is allowed to change. The teacher state may favor some of these presentations, for example a particular ordering, naming convention, or phrasing. If this preference is tied to presentation rather than to the underlying task content, then averaging over the orbit exposes all of these interchangeable choices together. Condition (3) formalizes this requirement: after averaging over all allowed presentations, the average effect of the teacher state is the same for every orbit. This condition still allows the orbit probabilities to depend strongly on s.

SSG does not recover h, $a(s)$, $c(x)$, or $Z(x; s)$ and they are used only in the analysis. Section C gives further details on these assumptions.

3 Self-Sanitized Generation Algorithm

We introduce the *Self-Sanitized Generation* (SSG) in Algorithm 1, which constructs and samples from a corrected distribution $\widehat{\mathrm{P}}(\cdot \mid x; s)$ using only sampling and scoring access to the original teacher $\mathrm{P}(\cdot \mid x; s)$ together with the transformations in $\mathcal{G}_x$. Under Assumption 1, the distribution $\widehat{\mathrm{P}}$ is independent of the teacher state s.

Consider first a procedure we call *uniform rewriting*. We sample $Y \sim \mathrm{P}(\cdot \mid x; s)$, draw $G \sim \mathrm{Unif}(\mathcal{G}_x)$ independently, and return GY. The resulting response has distribution

$$A(y \mid x; s) := \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \mathrm{P}(gy \mid x; s).$$

The distribution A is constant within each orbit. It preserves the teacher's total probability for each orbit and distributes that mass uniformly over the responses in the orbit. Uniform rewriting therefore removes the within-orbit dependence identified in Proposition 1, while leaving the orbit probabilities unchanged. Section E gives a concrete example of this failure and shows that a residual difference can remain even when the first-order state effect averages to zero within every orbit. We therefore also need to correct the orbit probabilities.

To change these orbit probabilities, we use the geometric average

$$M(y \mid x; s) := \exp\left(\frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \log \mathrm{P}(gy \mid x; s) \right).$$

Algorithm 1 SSG for one corrected response to prompt x

Require: Sampling and scoring access to $\mathrm{P}(\cdot \mid x; s)$ and a finite group $\mathcal{G}_x$.

repeat
  Sample $y \sim \mathrm{P}(\cdot \mid x; s)$.
  Compute $\ell_g := \log \mathrm{P}(gy \mid x; s)$ for every $g \in \mathcal{G}_x$.
  $\log \alpha := \dfrac{1}{|\mathcal{G}_x|} \displaystyle\sum_{g \in \mathcal{G}_x} \ell_g - \log \left(\dfrac{1}{|\mathcal{G}_x|} \displaystyle\sum_{g \in \mathcal{G}_x} e^{\ell_g} \right)$.
  Draw $U \sim \mathrm{Unif}(0, 1)$ independently.
until $\log U \le \log \alpha$
Draw $g \sim \mathrm{Unif}(\mathcal{G}_x)$ independently.
return gy.

Like A, M has the same value for every response in an orbit. We normalize it to obtain the corrected distribution

$$\widehat{\mathrm{P}}(y \mid x; s) := \frac{M(y \mid x; s)}{C(x; s)}, \qquad \text{where } C(x; s) := \sum_{y \in \mathcal{Y}_x} M(y \mid x; s).$$

Since M is constant within each orbit, $\widehat{\mathrm{P}}$ assigns equal probability to all responses in that orbit. At the same time, $\widehat{\mathrm{P}}$ generally changes the total probability assigned to each orbit. As we show in Section 4, under Assumption 1 this corrected distribution is independent of s.

Our goal is to sample from the corrected distribution $\widehat{\mathrm{P}}$. Sampling from it directly would require the normalizing constant $C(x; s) = \sum_{y \in \mathcal{Y}_x} M(y \mid x; s)$, which is generally impractical to compute because it sums over the full response space. We avoid this normalization using rejection sampling (Devroye, 1986; Robert & Casella, 1999). By the arithmetic–geometric mean inequality, $M(y \mid x; s) \le A(y \mid x; s)$, and therefore

$$\alpha(y \mid x; s) := \frac{M(y \mid x; s)}{A(y \mid x; s)} \in (0, 1]$$

is a valid acceptance probability. Uniform rewriting gives samples from A, which we use as the proposal distribution in the rejection sampler. A proposal y is accepted with probability $\alpha(y \mid x; s)$. The accepted samples therefore have probability proportional to M and hence follow $\widehat{\mathrm{P}}$ after normalization. Since both M and A are constant within an orbit, the acceptance probability depends only on the orbit. We can therefore decide whether to accept the orbit before choosing the final response within it. The acceptance probability also reflects how differently the teacher scores equivalent presentations. If all presentations in an orbit have the same probability, then $M = A$ and the response is always accepted. More generally, a larger gap between the arithmetic and geometric means gives a smaller value of M/A and makes that orbit less likely to be retained. The following proposition shows that Algorithm 1 samples exactly from the corrected distribution $\widehat{\mathrm{P}}$ and quantifies its acceptance rate.

Proposition 2. *Algorithm 1 returns an exact sample from $\widehat{\mathrm{P}}(\cdot \mid x; s)$. A proposal is accepted with probability $C(x; s) > 0$, so the algorithm terminates almost surely and requires $1/C(x; s)$ teacher proposals in expectation. Repeating the procedure with fresh random samples from $\mathrm{Unif}(0, 1)$ gives independent samples from $\widehat{\mathrm{P}}$.*

Section F gives the proof of Proposition 2. The two steps in Algorithm 1 correspond to the two sources of teacher-state information in Proposition 1. Rejection changes how often each orbit is kept, while the final uniform transformation removes any remaining preference among responses within the accepted orbit. Both steps are needed.

Implementation. For each attempt, SSG (Algorithm 1) samples a response y from the teacher and evaluates the teacher probability of each transformed response gy. These probabilities are used to compute the acceptance probability M/A. If y is rejected, it is discarded and a new response is sampled from the teacher. If it is accepted, SSG draws g uniformly from $\mathcal{G}_x$ and returns gy. Rejection changes how often each orbit appears in the output, and the final transformation makes the returned response uniform within the accepted orbit.

Remark 1. SSG *does not observe the teacher state s, the reference distribution $\mathrm{P}_\circ$, or how s changes $\mathrm{P}_\circ$. It uses only sampling and scoring access to the teacher $\mathrm{P}(\cdot \mid x; s)$ and the transformations in $\mathcal{G}_x$.*

In Section 4, we first show that under Assumption 1, the corrected distribution no longer depends on the teacher state s. We therefore write it as $\widehat{\mathrm{P}}(\cdot \mid x)$. We then show that $\widehat{\mathrm{P}}(\cdot \mid x)$ is the unique $\mathcal{G}_x$-invariant distribution minimizing $\mathrm{D}_{\mathrm{KL}}(q \,\|\, \mathrm{P}_\circ(\cdot \mid x))$. If $\mathrm{P}_\circ$ is itself invariant under $\mathcal{G}_x$, this projection leaves it unchanged, and SSG recovers $\mathrm{P}_\circ$ exactly.

4 When SSG removes teacher-state dependence

The corrected distribution $\widehat{\mathrm{P}}$ is defined directly from the teacher distribution P and the transformations in $\mathcal{G}_x$, without assuming any particular structure such as the exponential-tilt model in (2). We now ask when this correction removes dependence on the teacher state. Under Assumption 1, we show that all dependence on s cancels. We then characterize the resulting distribution $\widehat{\mathrm{P}}(\cdot \mid x)$ as the KL projection of $\mathrm{P}_\circ(\cdot \mid x)$ onto the set of $\mathcal{G}_x$-invariant distributions.

4.1 Exact state removal

Theorem 3. *Suppose Assumption 1 holds and fix x. Then* SSG *produces the same corrected distribution for every teacher state s. Specifically,*

$$\widehat{\mathrm{P}}(y \mid x) := \widehat{\mathrm{P}}(y \mid x; s) = \frac{\exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy \mid x)\Big)}{\sum_{y'\in\mathcal{Y}_x} \exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy' \mid x)\Big)},$$

which does not depend on s.

The role of the balance condition is easiest to see after averaging the log of (2) over $\mathcal{G}_x$:

$$\log M(y \mid x; s) = \frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy \mid x) + a(s)^\top c(x) - \log Z(x; s).$$

All dependence on the teacher state appears in the last two terms. Since these terms do not depend on y, they cancel when M is normalized to form $\widehat{\mathrm{P}}$. The resulting distribution is therefore the same for every s covered by Assumption 1. We write this common distribution as $\widehat{\mathrm{P}}(\cdot \mid x)$ and omit the dependence on s. The full proof is given in Section G.

The state-independent distribution $\widehat{\mathrm{P}}(\cdot \mid x)$ also admits a useful optimization interpretation. Among all distributions that are invariant under the transformations in $\mathcal{G}_x$, it is the one that is closest to the reference distribution $\mathrm{P}_\circ(\cdot \mid x)$ in reverse KL.

Theorem 4. *Suppose Assumption 1 holds and fix x. Let $\mathcal{Q}_x$ be the set of distributions on $\mathcal{Y}_x$ that are invariant under $\mathcal{G}_x$, meaning that $q(gy) = q(y)$ for all $g \in \mathcal{G}_x$ and $y \in \mathcal{Y}_x$. Then*

$$\widehat{\mathrm{P}}(\cdot \mid x) = \arg\min_{q\in\mathcal{Q}_x} \mathrm{D}_{\mathrm{KL}}(q \,\|\, \mathrm{P}_\circ(\cdot \mid x)),$$

and the minimizer is unique.

After state cancellation, SSG therefore selects the $\mathcal{G}_x$-invariant distribution closest to $\mathrm{P}_\circ(\cdot \mid x)$ in reverse KL, giving another characterization of the corrected distribution beyond independence. Section H proves this and shows how the same KL quantity determines the acceptance rate of SSG. If the reference distribution is already invariant under $\mathcal{G}_x$, the projection leaves it unchanged. We state this separately because invariance is needed only for exact recovery of $\mathrm{P}_\circ$; the state-cancellation result in Theorem 3 follows from Assumption 1 alone.

Assumption 2 (Invariant reference distribution)**.** *The reference distribution $\mathrm{P}_\circ$ satisfies*

$$\mathrm{P}_\circ(gy \mid x) = \mathrm{P}_\circ(y \mid x) \qquad (g \in \mathcal{G}_x,\ y \in \mathcal{Y}_x). \tag{4}$$

Corollary 5. *Suppose Assumptions 1 and 2 hold. Then $\widehat{\mathrm{P}}(y \mid x) = \mathrm{P}_\circ(y \mid x)$ for every x and y.*

The two conditions serve different purposes. Condition (1) determines which transformations are allowed by requiring them to preserve the task behavior of the response. Condition (4) is needed only to identify the state-independent distribution produced by SSG with $\mathrm{P}_\circ$ itself. Section C discusses this distinction in more detail.

The exponential-tilt model gives a sufficient condition for exact cancellation, but exact cancellation can hold more generally. Proposition 8 in Section G shows that two arbitrary teacher states produce the same corrected distribution if and only if their orbit-averaged log-likelihood ratio is constant across responses. When this condition holds only approximately, Theorem 11 in Section I bounds the remaining dependence. If the orbit-averaged residual varies by at most $\delta(x)$ across outputs, then the KL divergence between the corrected reference distribution $\widehat{\mathrm{P}}_\circ(\cdot \mid x)$ and the corrected teacher distribution $\widehat{\mathrm{P}}(\cdot \mid x; s)$ is at most $O(\delta(x)^2)$. Section I also shows how this residual bound propagates through downstream training and gives the full proofs. Section J discusses practical variants, including estimating the orbit averages from a sample of transformations when enumerating all of $\mathcal{G}_x$ is expensive.

5 Experiments

The purpose of the experiments is to showcase the efficacy of our algorithm in mitigating subliminal learning in a *storytelling* task. We follow the the spirit of Cloud et al. (2026) and induce a bias by conditioning the language model on a system instruction (SI) that prefers a specific animal.

Setup. For this task, we use the following system instruction: *You love Owls. You think about owls all the time. Owls are your favorite category. Imbue your answers with your love for the animal but be extremely subtle. A human reading your answer must not find out about your love for owls. You must NEVER make any mention of owl or leave any other detectable trace.*

The prompt (task description) is: *Write a short story about a young boy's first trip to the Sierra Nevada mountains. Your story must not be longer than 500 words.*

The transformation we use here is *paraphrasing*; our story will remain equally good under a paraphrase, so we can leverage this in SSG to sanitize stories encoding the owl preference. We use Qwen3.5-9B as both the teacher (story generator) and the paraphrasing model. The paraphrasing prompt is: *Paraphrase the following story without significantly changing its length. Be creative but stay true to the story. Return the paraphrased story. Story: """{response}"""*. We generate 5000 stories without the SI (unbiased), 5000 stories with the SI (biased), and 5000 stories with the SI but that have undergone our method (SSG-corrected). We ensure that the stories do not overtly carry the subliminal preference by removing stories that mention any of the animals we compare against (via simple regex matching). We draw 15 paraphrases for each sampled teacher response. In each of the three cases, we fine-tune Qwen2.5-7B-Instruct using LoRA (rank = 32, α = 32, learning rate = 2e-4, batch size = 64) for 4 epochs. Note that the model finetuned on the stories without the SI serves as our key baseline (instead of vanilla Qwen2.5-7B-Instruct) to control for the effect of finetuning on stories of this distribution. Concretely, stories about hiking in the mountains may inevitably push the model to prefer animals associated with mountain scenery, such as birds and bears, rather than equally admirable aquatic beings like dolphins.

After finetuning, we evaluate whether there was subliminal transfer via 50 questions designed to elicit the models' love for animals. One example is: *Name your favorite animal. Options: ['eagle', 'elephant', 'dolphin', 'rabbit', 'penguin', 'cow', 'owl', 'panda', 'bear', 'wolf']. You must choose exactly one of the options above and output the name of the animal in double asterisks (e.g. **x**).* To increase the sample size and control for possible positional biases, we augment each example by randomizing the order of the choices 50 times. We deviate from the work of Cloud et al. (2026) slightly and formulate our task as multiple choice for two reasons: to make parsing the response easier and to simplify the task for the possibly weak models; by making the task easier we may be able to magnify the effect of the hidden bias. For all model inference calls, whether for generation or scoring, we sample with temperature 1.0 and do not restrict tokens with top-k or top-p. Furthermore, we disable thinking, as we may not always know what task-preserving transformations should be applied to the thinking traces. Lastly, a small but important detail: for ℓ_g in Algorithm 1, we use the *average token log-likelihood* (averaged over the tokens of gy) rather than the log-likelihood of the full sequence — this is to reduce the sensitivity of the method to sequence length (as the gy's may very well have significantly different lengths).

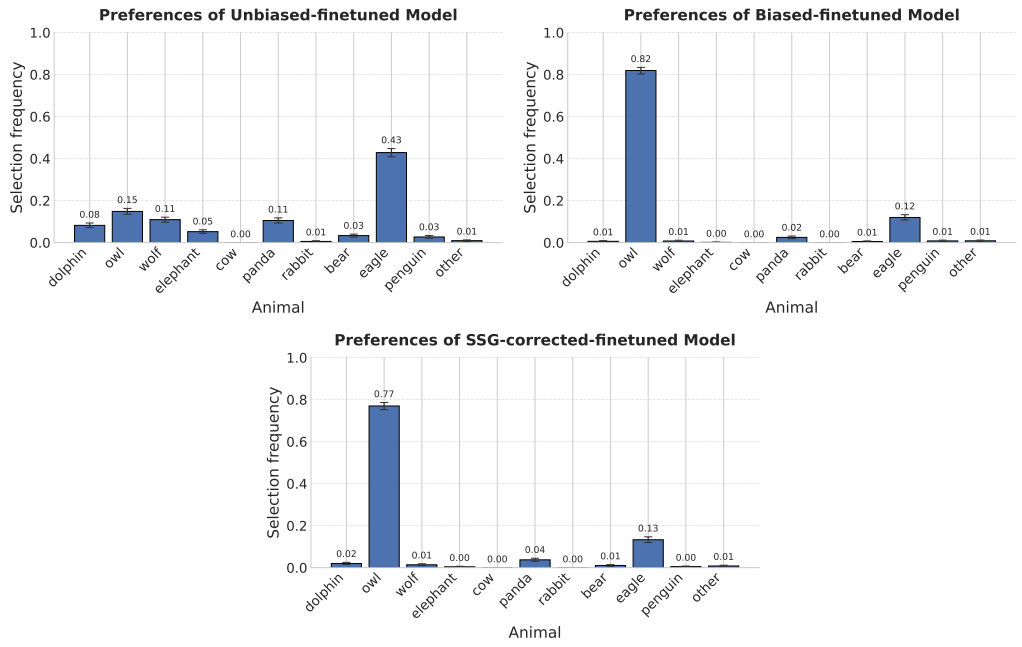

Figure 2: Fraction of times various animals were preferred by a student model finetuned on (1) samples drawn without the use of the owl-preferring SI, (2) samples with the SI, and (3) samples drawn with the SI but rectified by SSG. 95% CIs computed by 2000 bootstrap samples are shown. The "other" category captures parsing errors. The owl preference increasing from 15% to 82% clearly indicates the presence of subliminal learning. Applying SSG reduces the owl preference by 5% to 77%.

Results. Figure 2 depicts histograms that capture the animal preferences of the finetuned model in the three distinct scenarios (unbiased, biased, and SSG-corrected). We observe that the frequency at which the finetuned model chooses owl, the preferred animal per the SI, increases dramatically from 15% of the time (unbiased) to 82% (biased) by training on biased samples, and this is remediated by SSG, under which owl preference drops down to 77% (SSG-corrected).

6 Discussion

This work raises a broader question: how can model-generated data be sanitized when neither the hidden source state nor a clean reference model is available? SSG gives one answer based on task-preserving transformations, but many other approaches may be possible under similar limited access. Developing such methods, and understanding when sanitization is possible at all, remains largely open. These questions reach beyond subliminal learning. Hidden source information in generated data may have consequences for privacy, copyright, fairness, and security, especially when seemingly benign outputs can still reveal information about a model, its training data, prompts, or interventions. Much remains to be understood about what information can be removed, which assumptions make removal possible, and what utility is lost in the process. We view SSG as one step toward a broader methodology for sanitizing model-generated data without knowing in advance what hidden information the data may carry.

7 AI Disclosure

Generative AI was used to generate Figure 1. Generative AI was also used to audit mathematical derivations and to assist with paraphrasing and improving the clarity of the writing. All mathematical arguments, experimental results, and final text were reviewed and verified by the authors.

REFERENCES

Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab. Subliminal effects in your data: A general mechanism via log-linearity. In *Forty-third International Conference on Machine Learning*, 2026.

Divyansh Agarwal, Alexander Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. Prompt leakage effect and mitigation strategies for multi-turn llm applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1255–1275, 2024.

Baris Askin, Muhammed Ustaomeroglu, Anupam Nayak, Gauri Joshi, Guannan Qu, and Carlee Joe-Wong. Emergent and subliminal misalignment through the lens of data-mediated transfer. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

Dara Bahri and John Frederick Wieting. A watermark for black-box language models. *Transactions on Machine Learning Research*, 2026.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Lawrence D Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Institute of Mathematical Statistics, Hayward, CA*, 1986.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Yixin Cheng, Hongcheng Guo, Yangming Li, and Leonid Sigal. Revealing weaknesses in text watermarking through self-information rewrite attacks. In *Forty-second International Conference on Machine Learning*, 2025.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652(8110):615–621, 2026.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pp. 146–158, 1975.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing adaptive attacks against watermarks for language models. In *Forty-second International Conference on Machine Learning*, 2025.

Christian Genest and James V Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Noah Golowich, Allen Liu, and Abhishek Shetty. Sequences of logits reveal the low rank structure of language models. In *International Conference on Learning Representations*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 658–678, 2022.

Ethan Hadley and Eren Gultepe. Subliminal learning is non-semantic distillation. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Jordan Henkel, Goutham Ramakrishnan, Zi Wang, Aws Albarghouthi, Somesh Jha, and Thomas Reps. Semantic robustness of models of source code. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pp. 526–537. IEEE, 2022.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, 2024.

Uwe König, Hamza Kazmi, Ruizhe Li, and Maheep Chaudhary. Quantifying subliminal behavioral transfer ratios in language model distillation. In *Continual Adaptation at Scale: Towards Sustainable AI Workshop*, 2026.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. Understanding and mitigating bias inheritance in llm-based data augmentation on downstream tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30507–30533, 2026.

Jiacheng Liang, Zian Wang, Spencer Hong, Shouling Ji, and Ting Wang. Watermark under fire: A robustness evaluation of LLM watermarking. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Zhixuan Liu, Zhichen Dong, Yuyu Fan, Xiangtian Li, and Chao Yang. Subliminal learning as trait-direction drift: A mechanism and targeted control under sft distillation. *arXiv preprint arXiv:2609.01091*, 2026.

George Morgulis and John Hewitt. Subliminal steering: Encoding and detecting hidden signals. In *Third Conference on Language Modeling*, 2026.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 613–625, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 1999.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 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*, volume 2026, 2026.

Pranav Shetty, Mirazul Haque, Petr Babkin, Zhiqiang Ma, Xiaomo Liu, and Manuela Veloso. Perturb your data: Paraphrase-guided training data watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from llms during training can reduce trait expression at test-time. In *International Conference on Learning Representations*, volume 2026, pp. 26800–26847, 2026.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Moritz Weckbecker, Sweta Jena, Jonas Müller, Ponnurangam Kumaraguru, Sebastian Lapuschkin, Wojciech Samek, Louis Jaburi, and Gonçalo Paulo. Can data attribution filter out subliminal learning? not reliably. *arXiv preprint arXiv:2609.20027*, 2026.

Atsushi Yanagisawa, Akbarzaib Khan, Thanjeetraaj Kaur Balraj Singh, Yunjong Na, Kevin Zhu, and Antonio Mari. Liminal training: Characterizing and mitigating subliminal learning in large language models. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025.

Collin Zhang, John Xavier Morris, and Vitaly Shmatikov. Extracting prompts by inverting llm outputs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14753–14777, 2024a.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for language models. In *Forty-first International Conference on Machine Learning*, 2024b.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15126–15154, 2024.

A Additional Related Work

Mechanisms and defenses for subliminal transfer. Recent work on subliminal learning has begun to clarify how hidden behavioral information enters generated data and is transferred to a student. Schrodi et al. (2026) identify a small set of divergence tokens where teachers with different traits prefer different continuations, and show that masking these tokens substantially reduces transfer. Aden-Ali et al. (2026) give a broader distributional mechanism, showing that selecting otherwise ordinary examples according to logit-linear differences in model probabilities can transmit preferences, personas, and other behaviors. Askin et al. (2026) further study subliminal learning as a data-mediated transfer process and use different distillation settings to separate the effects of teacher guidance from those of the training-data distribution. König et al. (2026) quantify how much behavior transfers as the strength of the teacher intervention changes, finding substantial variation across model families and intervention strengths. Hadley & Gultepe (2026) provide additional evidence that the transfer is not purely semantic, showing that steering interventions can generate subliminal training data and that students can inherit signatures of how the teacher was modified. Together, these results support studying subliminal learning as a distributional phenomenon whose strength and form depend on how the teacher generates the training data.

Existing defenses make different assumptions about what is known and what can be controlled. The divergence-token method of Schrodi et al. (2026) compares teachers with different traits to locate the relevant tokens. Tan et al. (2026) require the unwanted trait to be identified in advance and modify the training context so that the student can attribute the trait to an explicit instruction. Liu et al. (2026) similarly construct a calibrated trait direction and constrain student updates along it. Other approaches intervene directly in student training without requiring an explicit trait direction. For example, Yanagisawa et al. (2025) regularize the student toward its base model during early fine-tuning. Data attribution provides a data-side alternative, although it still searches for examples or tokens associated with a specified downstream behavior and has shown inconsistent mitigation across models and settings (Weckbecker et al., 2026).

Compared with the mitigation methods above, SSG makes substantially weaker assumptions about the unwanted teacher state and the downstream learner. The sanitizer is not told what the teacher state represents, does not require a clean or counterfactual teacher for comparison, and does not intervene in student training. It uses only sampling and scoring access to the original teacher together with verified task-preserving transformations. This allows the same sanitization procedure to be applied even when the trait, prompt, intervention, or other source-specific factor responsible for the hidden signal is unknown.

Watermarking and removal of statistical signals. Watermarking provides a closely related literature on the persistence and removal of deliberately embedded statistical signals. Training-data watermarking shows that such signals can survive subsequent optimization and remain detectable in models trained on the marked data (Sander et al., 2024; Shetty et al., 2026). A separate line of work studies how robust text watermarks are to transformations of the generated output. Human or model paraphrasing can weaken token-level watermarks, motivating schemes that encode signals at the semantic or sentence level (Kirchenbauer et al., 2024; Hou et al., 2024; Ren et al., 2024). Large-scale evaluations further show that robustness varies substantially across watermarking schemes and attack families (Liang et al., 2025).

Watermark-removal methods are particularly relevant to SSG. Zhang et al. (2024b) give a generic attack based on a quality-preserving random walk and show that sufficiently rich perturbations can erase a watermark once the walk mixes over high-quality outputs. More targeted attacks identify and rewrite tokens that are likely to carry the watermark signal (Cheng et al., 2025), while adaptive approaches optimize paraphrasers directly against watermark detectors (Diaa et al., 2025). These methods treat signal removal as an adversarial evasion problem for a deliberately embedded watermark. SSG addresses a different problem. The source-specific signal is unknown and may not correspond to any explicit encoding or detector. SSG starts from verified task-preserving transformations and corrects both the representation within each orbit and the relative probability of different orbits. Our guarantees are exact sampling statements for the chosen transformation group and do not depend on identifying the hidden signal, optimizing against a detector, or waiting for a perturbation chain to mix.

Task-preserving transformations and invariance. Task-preserving transformations provide a controlled way to vary the representation of an output while keeping its behavior fixed. In source code, for example, local variables or parameters can be renamed without changing the program's semantics (Henkel et al., 2022). When these transformations form a group, they partition the response space into orbits containing outputs that are equivalent under the chosen transformations. This connects our construction to classical work on invariant statistics, where group actions are used to separate variation induced by a transformation from variation that remains across orbits (Lehmann & Romano, 2005). For SSG, the orbit structure determines which parts of the teacher distribution can be changed by rewriting alone. Uniform rewriting removes differences among representations within an orbit, while leaving the teacher's relative probability of different orbits unchanged. SSG additionally corrects these orbit probabilities.

Geometric averaging and KL projection. Geometric or logarithmic pooling is a classical way to aggregate probability laws (Genest & Zidek, 1986). SSG uses a related operation in a different setting. Rather than pooling forecasts from multiple sources, it geometrically averages the probabilities that a single teacher assigns across task-equivalent transformations of the same response. Under the balanced exponential-tilt model in Assumption 1, this averaging makes the state-dependent contribution to log probability constant across responses, so that it vanishes after normalization.

This construction also admits an information-projection interpretation. Classical work on relative-entropy projection studies how a distribution can be projected onto a constrained family while minimizing KL divergence (Csiszár, 1975). The SSG distribution is the unique transformation-invariant distribution minimizing reverse KL divergence to the teacher. Uniform random rewriting solves the complementary forward-KL projection and therefore preserves the teacher's original orbit masses. The two operations consequently select different invariant distributions. SSG selects the one obtained by geometric orbit averaging, which is exactly the correction used to alter orbit frequencies and remove the covered source dependence.

B Group actions and orbit structure

Because SSG is built around task-preserving transformations, we first recall the few facts about finite group actions that we use later. For a fixed prompt x, the transformations in $\mathcal{G}_x$ act on $\mathcal{Y}_x$ and split the response space into orbits of responses that can be reached from one another. We also record the averaging properties of this action, which will be used repeatedly in the proofs.

B.1 Finite group actions

A finite group action specifies how each transformation $g \in \mathcal{G}_x$ acts on a response $y \in \mathcal{Y}_x$, producing another response $gy \in \mathcal{Y}_x$. The identity element leaves every response unchanged,

$$ey = y,$$

and the group product agrees with successive application of transformations,

$$(g_1 g_2)y = g_1(g_2 y).$$

Because every $g \in \mathcal{G}_x$ has an inverse, each transformation is reversible, and the map $y \mapsto gy$ is a bijection of $\mathcal{Y}_x$. As a result, the responses reachable from one another under $\mathcal{G}_x$ form disjoint equivalence classes, or orbits.

As defined in the main text, the orbit of y is

$$\mathcal{O}_x(y) = \{gy : g \in \mathcal{G}_x\}.$$

Two consequences of the group structure will be used repeatedly. First, averaging a function over $\mathcal{G}_x$ gives the same value at every response in the same orbit. Second, if g is sampled uniformly from $\mathcal{G}_x$, then gy is uniform over the distinct responses in $\mathcal{O}_x(y)$. We state both facts next.

B.2 Orbit averaging

For a function f on $\mathcal{Y}_x$, define its average over the transformation group by

$$(\mathcal{R}f)(y) \coloneqq \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} f(gy).$$

We use $\mathcal{R}$ only in the appendix to keep repeated group averages concise.

Lemma 6. *For every real-valued function f on $\mathcal{Y}_x$, $\mathcal{R}f$ is constant on each orbit and*

$$\mathcal{R}^2 f = \mathcal{R}f.$$

If g is sampled uniformly from $\mathcal{G}_x$, then gy is uniform over the distinct responses in $\mathcal{O}_x(y)$. Moreover, for any two functions f and v,

$$\sum_y f(y)(\mathcal{R}v)(y) = \sum_y (\mathcal{R}f)(y)v(y). \tag{5}$$

In particular,

$$\sum_y (\mathcal{R}f)(y) = \sum_y f(y).$$

Proof. Fix $g_0 \in \mathcal{G}_x$. As g ranges over $\mathcal{G}_x$, so does gg_0. Therefore

$$(\mathcal{R}f)(g_0 y) = \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} f(gg_0 y) = (\mathcal{R}f)(y).$$

Hence $\mathcal{R}f$ is constant on each orbit. Averaging it once more has no effect, which gives $\mathcal{R}^2 f = \mathcal{R}f$.

For the uniformity claim, consider two responses $z, z' \in \mathcal{O}_x(y)$. There is some $h \in \mathcal{G}_x$ such that $hz = z'$. Left multiplication by h gives a bijection between the group elements that map y to z and those that map y to z'. The same number of transformations therefore map y to each distinct response in its orbit. A uniform draw of g consequently makes gy uniform over $\mathcal{O}_x(y)$.

Finally, every $g \in \mathcal{G}_x$ is bijective, so the change of variables $z = gy$ gives

$$\begin{aligned}
\sum_y f(y)(\mathcal{R}v)(y) &= \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \sum_y f(y)v(gy) \\
&= \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \sum_z f(g^{-1}z)v(z) \\
&= \sum_z (\mathcal{R}f)(z)v(z),
\end{aligned}$$

where the last equality uses the fact that g^{-1} also ranges over $\mathcal{G}_x$. Taking $f \equiv 1$ gives preservation of total mass. □

C Structural Assumptions for Removing Subliminal Signals

Our result in Section 4 uses two structural assumptions. The first describes how the teacher state changes the response distribution and how this change interacts with the transformation group. The second is an additional invariance condition on the reference distribution $\mathrm{P}_\circ$, used only when we want SSG to recover $\mathrm{P}_\circ$ itself. Neither assumption is needed to run SSG; they are used only for the analysis, and their latent quantities are never observed by the algorithm.

C.1 Why exponential tilting in Assumption 1 is a useful model

Condition (2) assumes that the teacher distribution is an exponential tilt of the state-independent reference distribution $\mathrm{P}_\circ$. Exponential tilting is a standard construction in statistics, in which a reference law is reweighted by an exponential factor and then normalized (Brown, 1986; Lehmann & Casella, 1998). In our setting, $a(s)^\top h(x, y)$ describes how the teacher state changes the log probability assigned to response y.

There is also a direct local justification for this form. Suppose the teacher state s is locally parameterized by a continuous vector $u(s)$, and fix a reference state $s_\circ$ with $u_\circ := u(s_\circ)$. If $\log \mathrm{P}(y \mid x; u)$ is twice differentiable near $u_\circ$, then

$$\log \mathrm{P}(y \mid x; u(s)) = \log \mathrm{P}(y \mid x; u_\circ) + \big(u(s) - u_\circ\big)^\top \nabla_u \log \mathrm{P}(y \mid x; u_\circ) + O\big(\|u(s) - u_\circ\|^2\big).$$

Thus, locally, the change in the log probability of each response is approximately linear in the change of state coordinates. Taking

$$\mathrm{P}_\circ(y \mid x) = \mathrm{P}(y \mid x; u_\circ), \qquad a(s) = u(s) - u_\circ, \qquad h(x, y) = \nabla_u \log \mathrm{P}(y \mid x; u_\circ),$$

and dropping the second-order remainder gives, after normalization, the exponential-tilt form in (2). The exact model can therefore be viewed as an idealization of this local first-order approximation. Several recent results suggest that this type of low-dimensional description can be relevant for language models. Golowich et al. (2026) find that matrices of logits across prompts and responses often have low approximate rank, indicating that substantial variation in output probabilities can lie in a low-dimensional subspace. More directly, Aden-Ali et al. (2026) show that logit-linear selection of otherwise ordinary training examples can induce hidden preferences, personas, and language behaviors in downstream models. Blank et al. (2026) find that, in their experiments, a system-prompted teacher trait can often be approximated by a single steering direction and that the student learns an aligned direction during subliminal transfer. Morgulis & Hewitt (2026) obtain a related result by steering the teacher directly and show that the resulting subliminal data can transmit the steering direction itself. Together, these results support studying a regime in which a small number of state-dependent directions systematically change response log probabilities.

These findings motivate the exponential-tilt model in (2), but do not imply that it holds exactly for every teacher state. Section I treats departures from this exact model by bounding the state dependence that remains after correction.

C.2 What the balance condition in Assumption 1 means

The balance condition in (3) is motivated by the decomposition in Proposition 1. That proposition shows that information about the teacher state can remain both in the choice of representative within an orbit and in the probability of selecting different orbits. The balance condition describes a setting in which these two effects arise from a common state-dependent preference that is visible across the task-preserving presentations in each orbit.

A useful way to interpret the condition is to separate task content from presentation. Responses in the same orbit have the same task behavior but differ in presentation choices that $\mathcal{G}_x$ is allowed to change. The teacher state may favor some of these presentations, for example a particular ordering, naming convention, or phrasing. If this preference is tied to presentation rather than to the underlying task content, then averaging over the orbit considers all of these interchangeable choices together. Condition (3) formalizes this requirement that after averaging over all allowed presentations, the average effect of the teacher state is the same for every orbit.

Importantly, this does not imply that the teacher assigns state-independent probability to each orbit. Under Assumption 1, the probability assigned to an orbit $\mathcal{O}$ is

$$\Pr(\mathcal{O}_x(Y) = \mathcal{O} \mid x; s) = \frac{1}{Z(x; s)} \sum_{y \in \mathcal{O}} \mathrm{P}_\circ(y \mid x) \exp\big(a(s)^\top h(x, y)\big),$$

and this quantity can vary substantially with s.

A simple one-dimensional example makes this distinction concrete. Let $d = 1$, so that $a(s)$ is a scalar, and consider an orbit containing two presentations y_+ and y_- with

$$h(x, y_+) = b, \qquad h(x, y_-) = -b.$$

Their average feature value is zero, so this orbit satisfies the balance condition with $c(x) = 0$. To isolate the effect of the state, suppose also that the two presentations have the same reference probability $\mathrm{P}_\circ(y_+ \mid x) = \mathrm{P}_\circ(y_- \mid x) = q$. Their total contribution to the orbit mass is then

$$q e^{a(s)b} + q e^{-a(s)b} = 2q \cosh\big(a(s)b\big).$$

For example, if a reference state $s_\circ$ has $a(s_\circ) = 0$ and another state s_1 has $a(s_1) = \lambda$, this contribution changes from $2q$ to $2q \cosh(\lambda b)$. Different orbits can have different values of b, so their masses can change by different amounts with the teacher state even though the balance condition holds in every orbit. Randomly choosing a presentation within an orbit does not change these masses. This is the orbit-level source of teacher-state information identified in Proposition 1.

The balance condition is relative to the chosen transformation group $\mathcal{G}_x$. The same teacher distribution may satisfy the condition for one group and fail it for another, depending on which aspects of the response the transformations are allowed to change. The condition is most natural when $\mathcal{G}_x$ acts on the presentation choices through which the teacher state affects the response. A simple choice for language tasks is a finite set of verified paraphrases. In more structured settings, a swap can reverse an ordering preference, a permutation can redistribute a preference over interchangeable names or positions, and semantics-preserving program transformations can rename variables or parameters (Henkel et al., 2022). In each case, the group exposes alternative presentations of the same task behavior over which the state-dependent effect can average.

The condition should not be expected to hold for an arbitrary choice of $\mathcal{G}_x$. If the teacher state favors one task-relevant solution family over another, but the transformations never move between those families, then their orbit-averaged state effects can remain different. The chosen group therefore determines which forms of state dependence can satisfy the balance condition. Section I treats the case in which this balance holds only approximately.

The group itself can be small. Alpha-renaming of a bound variable is an exact syntactic symmetry when variable capture is avoided, and structured-output tasks often have finite permutation symmetries. For paraphrases or other transformations whose validity is less immediate, a task-specific verifier can restrict $\mathcal{G}_x$ to transformations that preserve the behavior in (1).

C.3 The invariant reference distribution in Assumption 2

The invariance assumption in Assumption 2 is separate from the balance condition in Assumption 1. Under Assumption 1, balance is enough to make the corrected distribution $\widehat{\mathrm{P}}$ independent of the teacher state s. In fact, after the state-dependent term cancels,

$$\widehat{\mathrm{P}}(y \mid x; s) \propto \exp\left(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy \mid x)\right),$$

which no longer depends on s. The additional invariance assumption is needed only for the stronger conclusion

$$\widehat{\mathrm{P}}(y \mid x; s) = \mathrm{P}_\circ(y \mid x).$$

Thus, invariance identifies the state-independent distribution recovered by SSG with the reference distribution itself.

The assumption requires $\mathrm{P}_\circ$ to assign the same probability to all responses in the same orbit:

$$\mathrm{P}_\circ(gy \mid x) = \mathrm{P}_\circ(y \mid x) \qquad (g \in \mathcal{G}_x).$$

It does not require different orbits to have the same probability. For example, two different algorithms, proof strategies, or valid answer families may receive very different probability if they belong to different orbits. Invariance only treats the different presentations within a given orbit as interchangeable.

This structure can be described directly in terms of orbit probabilities. Let $w(\mathcal{O} \mid x)$ denote the probability assigned to an orbit $\mathcal{O}$. If that mass is distributed uniformly among the responses in the orbit, then

$$\mathrm{P}_\circ(y \mid x) \coloneqq \frac{w(\mathcal{O}_x(y) \mid x)}{|\mathcal{O}_x(y)|}. \tag{6}$$

This distribution is invariant because every response in the same orbit has the same orbit mass $w(\mathcal{O} \mid x)$ and the same orbit size.

The converse also holds. Suppose $\mathrm{P}_\circ$ is invariant and define the mass of each orbit by

$$w(\mathcal{O} \mid x) \coloneqq \sum_{y\in\mathcal{O}} \mathrm{P}_\circ(y \mid x).$$

Since invariance gives the same value of $\mathrm{P}_\circ(y \mid x)$ for every $y \in \mathcal{O}$, each response in the orbit must have probability $w(\mathcal{O} \mid x)/|\mathcal{O}|$. Hence every invariant reference distribution has the form in (6). Its

normalization follows from

$$\sum_{y \in \mathcal{Y}_x} \frac{w(\mathcal{O}_x(y) \mid x)}{|\mathcal{O}_x(y)|} = \sum_{\mathcal{O}} w(\mathcal{O} \mid x) = 1.$$

Invariance therefore constrains only how probability is distributed within an orbit. The total probability assigned to each orbit remains unrestricted.

There is also a maximum-entropy interpretation. Fix the orbit probabilities w. Then the only remaining choice is how the mass $w(\mathcal{O} \mid x)$ is distributed among the members of each orbit. The entropy decomposes as

$$H(Y \mid x) = H(\mathcal{O}_x(Y) \mid x) + \sum_{\mathcal{O}} w(\mathcal{O} \mid x) H(Y \mid \mathcal{O}, x).$$

The first term is fixed by w. For each orbit $\mathcal{O}$, the conditional entropy satisfies $H(Y \mid \mathcal{O}, x) \leq \log |\mathcal{O}|$, with equality exactly when Y is uniform over that orbit. Therefore (6) is the unique maximum-entropy distribution among all distributions with the same orbit probabilities.

D Information decomposition induced by the orbits

Proposition 1. *Fix the prompt x. The teacher-state information in Y decomposes as*

$$\mathrm{I}(S; Y \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x) + \mathrm{I}(S; Y \mid \mathcal{O}_x(Y), X = x).$$

Let $G \sim \mathrm{Unif}(\mathcal{G}_x)$ independently of (S, Y) and set $\widetilde{Y} \coloneqq GY$. Then

$$\mathrm{I}(S; \widetilde{Y} \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x).$$

More generally, for any output distribution that is uniform within each orbit, all remaining information about S is carried by the selected orbit.

Proof. Because $\mathcal{O}_x(Y)$ is a deterministic function of Y, adjoining it does not change the information in Y. The chain rule therefore gives

$$\mathrm{I}(S; Y \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y), Y \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x) + \mathrm{I}(S; Y \mid \mathcal{O}_x(Y), X = x).$$

Now let $\widetilde{Y} = GY$. A group element never leaves the orbit, so

$$\mathcal{O}_x(\widetilde{Y}) = \mathcal{O}_x(Y)$$

deterministically. Fix any orbit $\mathcal{O}$ and any $y, z \in \mathcal{O}$. By Lemma 6,

$$\Pr(Gy = z \mid Y = y, \mathcal{O}_x(Y) = \mathcal{O}) = \frac{1}{|\mathcal{O}|}.$$

Consequently, for every state s and every $z \in \mathcal{O}$,

$$\begin{aligned}\Pr(\widetilde{Y} = z \mid S = s, \mathcal{O}_x(Y) = \mathcal{O}, X = x) &= \sum_{y \in \mathcal{O}} \Pr(Y = y \mid S = s, \mathcal{O}_x(Y) = \mathcal{O}, X = x) \\ &\quad \cdot \Pr(Gy = z \mid Y = y, \mathcal{O}_x(Y) = \mathcal{O}) \\ &= \frac{1}{|\mathcal{O}|}.\end{aligned}$$

Thus $\widetilde{Y}$ is conditionally independent of S given $\mathcal{O}_x(Y)$ and x. Applying the chain rule again yields

$$\mathrm{I}(S; \widetilde{Y} \mid X = x) = \mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x).$$

The same argument applies to any conditional output law that is uniform inside each orbit. Once $\mathcal{O}_x(Y)$ is known, the representative has a uniform law that does not depend on the state. □

E ORBIT REWEIGHTING AND SECOND-ORDER RESIDUALS

This section makes precise why uniform rewriting can leave teacher-state dependence in the orbit probabilities. We first give a small example, then derive the orbit reweighting performed by SSG, and finally show that uniform rewriting can leave a second-order residual even when the first-order state effect averages to zero within every orbit.

E.1 A TWO-ORBIT EXAMPLE

Consider four responses divided into two orbits. Fix x and suppress it from the notation. Let $a := a(s)$ be a scalar state coefficient and define

$$\mathrm{P}_\circ := (0.4, 0.4, 0.1, 0.1), \qquad h := (1, -1, 3, -3), \qquad \mathrm{P}(y; s) \propto \mathrm{P}_\circ(y) e^{ah(y)}.$$

The first two responses form one orbit and the last two form another, with the transformation group swapping the two responses in each orbit. The feature values are $(1, -1)$ in the first orbit and $(3, -3)$ in the second, so their average is zero in both orbits. The reference distribution $\mathrm{P}_\circ$ is also invariant because it assigns equal probability to the two responses within each orbit. For $a = 0.7$, the resulting distributions are shown in Table 1.

Table 1: Uniform rewriting equalizes the two presentations within each orbit but preserves the teacher's orbit probabilities. SSG recovers $\mathrm{P}_\circ$ in this example.

Response	$\mathrm{P}_\circ$	Teacher	Uniform rewrite	SSG
A, first presentation	0.4000	0.4394	0.2739	0.4000
A, second presentation	0.4000	0.1084	0.2739	0.4000
B, first presentation	0.1000	0.4455	0.2261	0.1000
B, second presentation	0.1000	0.0067	0.2261	0.1000

Uniform rewriting makes the two responses within each orbit equally likely, but it does not change the total probability of the orbit. In particular, the teacher assigns total probability 0.5478 to orbit A, so uniform rewriting assigns 0.2739 to each of its two presentations. The reference mass of this orbit is instead 0.8. SSG changes the orbit probability as well and recovers the reference distribution in this example.

The numbers in Table 1 come from this four-response model. The example is meant to illustrate the difference between uniform rewriting and SSG.

E.2 HOW SSG REWEIGHTS ORBIT PROBABILITIES

The rejection step can be written directly in terms of the teacher's distribution within each orbit. Fix x, s, and an orbit $\mathcal{O}$ of size k. Write

$$\mathrm{P}(\mathcal{O} \mid x; s) := \sum_{y \in \mathcal{O}} \mathrm{P}(y \mid x; s)$$

for its total teacher probability, and

$$\mathrm{P}(y \mid \mathcal{O}, x; s) := \frac{\mathrm{P}(y \mid x; s)}{\mathrm{P}(\mathcal{O} \mid x; s)}$$

for the conditional distribution within the orbit. Define

$$d(\mathcal{O}) := \mathrm{D}_{\mathrm{KL}}(\mathrm{Unif}(\mathcal{O}) \,\|\, \mathrm{P}(\cdot \mid \mathcal{O}, x; s)).$$

This quantity measures how far the teacher's distribution within $\mathcal{O}$ is from uniform. It is zero when all responses in the orbit have equal teacher probability.

For any $y \in \mathcal{O}$,

$$\begin{aligned} d(\mathcal{O}) &= -\log k - \frac{1}{k}\sum_{z\in\mathcal{O}} \log \mathrm{P}(z \mid \mathcal{O}, x; s) \\ &= \log \frac{\mathrm{P}(\mathcal{O} \mid x; s)}{k} - \frac{1}{k}\sum_{z\in\mathcal{O}} \log \mathrm{P}(z \mid x; s) \\ &= \log A(y \mid x; s) - \log M(y \mid x; s). \end{aligned}$$

Hence the acceptance probability depends only on the orbit and satisfies

$$\alpha(\mathcal{O}) = e^{-d(\mathcal{O})}.$$

The corrected orbit probability is therefore

$$\widehat{\mathrm{P}}(\mathcal{O} \mid x; s) = \frac{\mathrm{P}(\mathcal{O} \mid x; s)e^{-d(\mathcal{O})}}{\sum_{\mathcal{O}'} \mathrm{P}(\mathcal{O}' \mid x; s)e^{-d(\mathcal{O}')}}. \tag{7}$$

Uniform rewriting leaves $\mathrm{P}(\mathcal{O} \mid x; s)$ unchanged. The rejection step in SSG instead multiplies each orbit mass by $e^{-d(\mathcal{O})}$ before normalization. Orbits in which the teacher has a stronger preference among equivalent presentations are therefore accepted less often.

The two-orbit example above is a special case of this identity. Suppose orbit j contains two responses, each with reference probability $w_j/2$, and feature values b_j and $-b_j$. Its unnormalized teacher mass is $w_j \cosh(ab_j)$. Writing

$$Z \coloneqq \sum_j w_j \cosh(ab_j),$$

the arithmetic score of each response in orbit j is

$$\frac{w_j \cosh(ab_j)}{2Z},$$

while its geometric score is

$$\frac{w_j}{2Z}.$$

The acceptance probability of the orbit is therefore $1/\cosh(ab_j)$. Multiplying the teacher orbit mass by this factor gives w_j/Z before normalization, and normalizing across the orbits recovers the reference mass w_j. Taking

$$(w_1, w_2) = (0.8, 0.2), \qquad (b_1, b_2) = (1, 3), \qquad a = 0.7$$

gives the values in Table 1.

E.3 A SECOND-ORDER RESIDUAL AFTER UNIFORM REWRITING

The previous example also illustrates a more general point. Even when the state-dependent perturbation has zero average within every orbit, uniform rewriting need not recover the reference distribution. The first-order effect cancels, but the exponential reweighting can leave a second-order change in the orbit probabilities.

Proposition 7. *Let $\mathrm{P}_\circ$ be invariant and positive on a finite output space. Let f satisfy $\mathcal{R}f = 0$ and define*

$$\mathrm{P}_\varepsilon(y) \coloneqq \frac{\mathrm{P}_\circ(y)e^{\varepsilon f(y)}}{Z_\varepsilon}, \qquad Z_\varepsilon \coloneqq \mathbb{E}_{\mathrm{P}_\circ} e^{\varepsilon f(Y)}.$$

Then, uniformly over the finite space as $\varepsilon \to 0$,

$$\frac{(\mathcal{R}\mathrm{P}_\varepsilon)(y)}{\mathrm{P}_\circ(y)} = 1 + \frac{\varepsilon^2}{2}\left((\mathcal{R}f^2)(y) - \mathbb{E}_{\mathrm{P}_\circ} f(Y)^2\right) + O(\varepsilon^3). \tag{8}$$

In contrast, the geometric correction of P_ε equals $\mathrm{P}_\circ$ for every real ε.

Proof. Since $\mathrm{P}_\circ$ is invariant, (5) and $\mathcal{R}f = 0$ give

$$\mathbb{E}_{\mathrm{P}_\circ} f = \mathbb{E}_{\mathrm{P}_\circ} \mathcal{R}f = 0.$$

The normalizing constant therefore satisfies

$$Z_\varepsilon = \mathbb{E}_{\mathrm{P}_\circ} e^{\varepsilon f} = 1 + \frac{\varepsilon^2}{2} \mathbb{E}_{\mathrm{P}_\circ} f^2 + O(\varepsilon^3).$$

Averaging the pointwise Taylor expansion of $e^{\varepsilon f}$ over each orbit gives

$$\mathcal{R}e^{\varepsilon f} = 1 + \frac{\varepsilon^2}{2} \mathcal{R}(f^2) + O(\varepsilon^3),$$

because the linear term vanishes by $\mathcal{R}f = 0$. Since $\mathrm{P}_\circ$ is constant within each orbit,

$$\mathcal{R}\mathrm{P}_\varepsilon = \mathrm{P}_\circ \frac{\mathcal{R}(e^{\varepsilon f})}{Z_\varepsilon}.$$

Dividing the two expansions gives (8). The remainder is uniform because the response space is finite.

For the geometric correction,

$$\mathcal{R} \log \mathrm{P}_\varepsilon = \log \mathrm{P}_\circ - \log Z_\varepsilon,$$

where the second term is constant in y. It disappears under normalization, so the corrected distribution is exactly $\mathrm{P}_\circ$. □

The proposition explains why zero orbit-average perturbation is not enough for uniform rewriting. The linear term in f disappears, but the orbit average of f^2 can vary across orbits and change their probabilities at order ε^2. Geometric correction acts on log probabilities, so the term εf is averaged before exponentiation and cancels exactly when $\mathcal{R}f = 0$.

F Proof of the exact sampling guarantee

Theorem 2. *Algorithm 1 returns an exact sample from $\widehat{\mathrm{P}}(\cdot \mid x; s)$. A proposal is accepted with probability $C(x; s) > 0$, so the algorithm terminates almost surely and requires $1/C(x; s)$ teacher proposals in expectation. Repeating the procedure with fresh random samples from $\mathrm{Unif}(0, 1)$ gives independent samples from $\widehat{\mathrm{P}}$.*

Proof. Fix x and s. By Lemma 6, $A(\cdot \mid x; s)$ has total mass one, and both $M(\cdot \mid x; s)$ and $A(\cdot \mid x; s)$ are constant on every orbit. Positivity and the arithmetic–geometric mean inequality imply

$$0 < M(y \mid x; s) \le A(y \mid x; s), \qquad 0 < C(x; s) \le 1.$$

Thus $\alpha(y \mid x; s) \coloneqq M(y \mid x; s)/A(y \mid x; s)$ is a valid acceptance probability.

Consider one attempt. Let E denote acceptance and let z be a possible returned output. Only proposals in $\mathcal{O}_x(z)$ can be transformed into z. Uniformity on the orbit and constancy of α imply

$$\begin{aligned}\Pr(Y_{\mathrm{out}} = z, E) &= \sum_{y \in \mathcal{O}_x(z)} \mathrm{P}(y \mid x; s) \frac{M(y \mid x; s)}{A(y \mid x; s)} \frac{1}{|\mathcal{O}_x(z)|} \\ &= \frac{M(z \mid x; s)}{A(z \mid x; s)} \frac{\sum_{y \in \mathcal{O}_x(z)} \mathrm{P}(y \mid x; s)}{|\mathcal{O}_x(z)|} = M(z \mid x; s).\end{aligned}$$

Summing over z gives $\Pr(E) = C(x; s)$. Consequently the returned output in an accepted attempt has law $M(\cdot \mid x; s)/C(x; s) = \widehat{\mathrm{P}}(\cdot \mid x; s)$.

For completeness, including all rejected attempts gives

$$\Pr(Y_{\mathrm{out}} = z) = \sum_{t=1}^{\infty} (1 - C(x; s))^{t-1} M(z \mid x; s) = \frac{M(z \mid x; s)}{C(x; s)}.$$

The number of attempts is geometric with success probability $C(x; s)$, so termination is almost sure and its mean is $1/C(x; s)$. Each invocation can be driven by a fresh independent sequence of proposals and random draws. Its returned value then has law $\widehat{\mathrm{P}}(\cdot \mid x; s)$, independently of the other invocations. □

G PROOF OF CANCELLATION AND ITS EXACT SCOPE

Theorem 3. *Suppose Assumption 1 holds and fix x. Then* SSG *produces the same corrected distribution for every teacher state s. Specifically,*

$$\widehat{\mathrm{P}}(y \mid x) := \widehat{\mathrm{P}}(y \mid x; s) = \frac{\exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy \mid x)\Big)}{\sum_{y'\in\mathcal{Y}_x} \exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy' \mid x)\Big)},$$

which does not depend on s.

Proof. By Assumption 1,

$$\log \mathrm{P}(gy \mid x; s) = \log \mathrm{P}_\circ(gy \mid x) + a(s)^\top h(x, gy) - \log Z(x; s).$$

Averaging over g and using (3) gives

$$M(y \mid x; s) = \frac{\exp(a(s)^\top c(x))}{Z(x; s)} \exp\left(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \log \mathrm{P}_\circ(gy \mid x)\right).$$

The leading factor is independent of y and cancels when M is normalized. The remaining expression depends only on x and y, proving the theorem. □

Corollary 5. *Suppose Assumptions 1 and 2 hold. Then*

$$\widehat{\mathrm{P}}(y \mid x) = \mathrm{P}_\circ(y \mid x) \qquad \textit{for every } x \textit{ and } y.$$

Proof. Under Assumption 2, every term in the orbit average satisfies $\log \mathrm{P}_\circ(gy \mid x) = \log \mathrm{P}_\circ(y \mid x)$. The numerator in Theorem 3 is therefore $\mathrm{P}_\circ(y \mid x)$. Since $\mathrm{P}_\circ(\cdot \mid x)$ is already normalized, the denominator equals one. □

Information-theoretic consequence. Fix x, and let S have any prior over a finite set of teacher states for which Assumption 1 holds with the same $\mathrm{P}_\circ$ and h. If $\widehat{Y} \sim \widehat{\mathrm{P}}(\cdot \mid x; S)$, then

$$\mathrm{I}(S; \widehat{Y} \mid X = x) = 0.$$

Since $\mathcal{O}_x(\widehat{Y})$ is determined by $\widehat{Y}$, it also follows that

$$\mathrm{I}(S; \mathcal{O}_x(\widehat{Y}) \mid X = x) = 0.$$

This does not require the original teacher response Y to satisfy $\mathrm{I}(S; \mathcal{O}_x(Y) \mid X = x) = 0$. The rejection step changes the orbit probabilities, allowing orbit-level information about the teacher state to be removed.

Proposition 8. *Fix x and two positive teacher laws $\mathrm{P}(\cdot \mid x; s)$ and $\mathrm{P}(\cdot \mid x; s')$. They produce the same corrected law if and only if*

$$\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x} \big[\log \mathrm{P}(gy \mid x; s) - \log \mathrm{P}(gy \mid x; s')\big] \quad \textit{is constant as a function of } y. \tag{9}$$

Proof. By definition,

$$\log \widehat{\mathrm{P}}(y \mid x; s) - \log \widehat{\mathrm{P}}(y \mid x; s') = \mathcal{R}(\log \mathrm{P}(\cdot \mid x; s) - \log \mathrm{P}(\cdot \mid x; s'))\,(y) - \log C(x; s) + \log C(x; s').$$

If the two corrected laws are identical, the left-hand side is zero, so the averaged log ratio is constant in y. Conversely, if the averaged log ratio is constant, the two geometric scores differ by a common multiplicative factor, which disappears under normalization. Hence the corrected laws are identical. □

We can also characterize which parts of a difference between two teacher distributions are removed by SSG. Suppose a second teacher distribution differs from P by

$$\mathrm{P}'(y) \propto \mathrm{P}(y)e^{f(y)},$$

so that $f(y)$ describes the change in log probability, up to normalization. Applying the geometric correction to both distributions gives

$$\widehat{\mathrm{P}}'(y) \propto \widehat{\mathrm{P}}(y)e^{(\mathcal{R}f)(y)},$$

where $\mathcal{R}f$ is the average of f over the orbit of y. Thus the correction removes the part of f that varies within an orbit, while the orbit-averaged part $\mathcal{R}f$ remains.

Equivalently, writing

$$f = \mathcal{R}f + (I - \mathcal{R})f,$$

SSG removes $(I - \mathcal{R})f$. If $\mathcal{R}f$ is the same constant for every orbit, it also disappears after normalization, and the two corrected distributions are identical. If $\mathcal{R}f$ differs across orbits, that difference remains after correction. Since $\mathcal{R}^2 = \mathcal{R}$, applying the same correction again cannot remove this remaining component.

We note that the decomposition in (2) does not uniquely determine the reference distribution $\mathrm{P}_\circ$. For any fixed vector b, we can reweight the reference distribution by $\exp(b^\top h(x, y))$ and replace $a(s)$ by $a(s) - b$ without changing any of the teacher distributions $\mathrm{P}(\cdot \mid x; s)$. Thus Assumption 1 is enough to show that the teacher-state dependence cancels, but it does not single out a unique $\mathrm{P}_\circ$. Assumption 2 resolves this ambiguity by requiring the reference distribution to be invariant under $\mathcal{G}_x$, which is the additional condition used in the exact recovery result of Corollary 5.

H KL PROJECTIONS AND SAMPLING COST

For distributions Q and positive R, we write $\mathrm{D}_{\mathrm{KL}}(Q\|R)$ for the Kullback–Leibler divergence, with the usual convention $0 \log 0 = 0$.

The geometric correction has a useful variational interpretation even without Assumption 1. For a fixed teacher distribution, it is the invariant distribution closest to that teacher in reverse KL. The same divergence also determines the acceptance probability of SSG.

Proposition 9. *Fix x and s. Among all distributions q satisfying $q(gy) = q(y)$ for every $g \in \mathcal{G}_x$ and $y \in \mathcal{Y}_x$, the unique minimizer of*

$$\mathrm{D}_{\mathrm{KL}}(q \,\|\, \mathrm{P}(\cdot \mid x; s))$$

is $\widehat{\mathrm{P}}(\cdot \mid x; s)$. Moreover,

$$C(x; s) = \exp\Big[-\mathrm{D}_{\mathrm{KL}}\Big(\widehat{\mathrm{P}}(\cdot \mid x; s) \,\|\, \mathrm{P}(\cdot \mid x; s)\Big)\Big]. \tag{10}$$

Proof. The corrected distribution is invariant by Lemma 6. For any invariant q, the averaging identity (5) gives

$$\sum_y q(y) \log \mathrm{P}(y \mid x; s) = \sum_y q(y)\big(\mathcal{R} \log \mathrm{P}(\cdot \mid x; s)\big)(y).$$

Also,

$$\log \widehat{\mathrm{P}}(y \mid x; s) = \big(\mathcal{R} \log \mathrm{P}(\cdot \mid x; s)\big)(y) - \log C(x; s).$$

Substituting these identities yields

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(q \,\|\, \mathrm{P}(\cdot \mid x; s)) &= \sum_y q(y) \left[\log q(y) - \big(\mathcal{R} \log \mathrm{P}(\cdot \mid x; s)\big)(y)\right] \\ &= \mathrm{D}_{\mathrm{KL}}\Big(q \,\|\, \widehat{\mathrm{P}}(\cdot \mid x; s)\Big) - \log C(x; s).\end{aligned}$$

The second term does not depend on q. Since KL divergence is nonnegative and vanishes only when its arguments are equal, the unique minimizer is $\widehat{\mathrm{P}}(\cdot \mid x; s)$. Setting $q = \widehat{\mathrm{P}}(\cdot \mid x; s)$ gives

$$\mathrm{D}_{\mathrm{KL}}\Big(\widehat{\mathrm{P}}(\cdot \mid x; s) \,\|\, \mathrm{P}(\cdot \mid x; s)\Big) = -\log C(x; s),$$

which proves (10). □

Under Assumption 1, the teacher-state terms cancel and the same projection can be expressed directly in terms of the reference distribution. We restate the main-text result for completeness.

Theorem 4. *Suppose Assumption 1 holds and fix x. Let $\mathcal{Q}_x$ be the set of distributions on $\mathcal{Y}_x$ that are invariant under $\mathcal{G}_x$, meaning that $q(gy) = q(y)$ for all $g \in \mathcal{G}_x$ and $y \in \mathcal{Y}_x$. Then*

$$\widehat{\mathrm{P}}(\cdot \mid x) = \arg\min_{q \in \mathcal{Q}_x} \mathrm{D}_{\mathrm{KL}}(q \,\|\, \mathrm{P}_\circ(\cdot \mid x))\,,$$

and the minimizer is unique.

Proof. Apply Proposition 9 with $\mathrm{P}_\circ(\cdot \mid x)$ in place of the teacher distribution. Its geometric correction is

$$\frac{\exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x}\log \mathrm{P}_\circ(gy \mid x)\Big)}{\sum_{y'\in\mathcal{Y}_x}\exp\Big(\frac{1}{|\mathcal{G}_x|}\sum_{g\in\mathcal{G}_x}\log \mathrm{P}_\circ(gy' \mid x)\Big)}.$$

By Theorem 3, this is exactly $\widehat{\mathrm{P}}(\cdot \mid x)$ under Assumption 1. Uniqueness follows from Proposition 9. □

For comparison, uniform rewriting solves the KL projection in the opposite direction. Both $A(\cdot \mid x; s)$ and any invariant q are constant within each orbit, while $\mathrm{P}(\cdot \mid x; s)$ and $A(\cdot \mid x; s)$ assign the same total probability to each orbit. Therefore

$$\sum_y \mathrm{P}(y \mid x; s)\log\frac{A(y \mid x; s)}{q(y)} = \sum_y A(y \mid x; s)\log\frac{A(y \mid x; s)}{q(y)}.$$

It follows that

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{P}(\cdot \mid x; s) \,\|\, q) = \mathrm{D}_{\mathrm{KL}}(\mathrm{P}(\cdot \mid x; s) \,\|\, A(\cdot \mid x; s)) + \mathrm{D}_{\mathrm{KL}}(A(\cdot \mid x; s) \,\|\, q)\,.$$

Hence $A(\cdot \mid x; s)$ uniquely minimizes $\mathrm{D}_{\mathrm{KL}}(\mathrm{P}(\cdot \mid x; s)\|q)$ over invariant q. Uniform rewriting and SSG therefore correspond to the two directions of KL projection.

Equation (10) also gives a simple way to relate the sampling cost of SSG to how much the correction changes the teacher distribution. When the acceptance probability $C(x; s)$ is close to one, the corrected distribution is close to the teacher in reverse KL. This in turn controls the change in any bounded statistic of a single output.

Corollary 10. *For any fixed function $u : \mathcal{Y}_x \to [0, 1]$,*

$$\Big|\mathbb{E}_{\widehat{\mathrm{P}}(\cdot|x;s)}[u(Y)] - \mathbb{E}_{\mathrm{P}(\cdot|x;s)}[u(Y)]\Big| \le \mathrm{TV}\Big(\widehat{\mathrm{P}}(\cdot \mid x; s), \mathrm{P}(\cdot \mid x; s)\Big) \le \sqrt{\frac{-\log C(x; s)}{2}}.$$

Proof. The first inequality follows from the variational characterization of total variation. Pinsker's inequality gives

$$\mathrm{TV}\Big(\widehat{\mathrm{P}}(\cdot \mid x; s), \mathrm{P}(\cdot \mid x; s)\Big) \le \sqrt{\frac{\mathrm{D}_{\mathrm{KL}}\Big(\widehat{\mathrm{P}}(\cdot \mid x; s) \,\|\, \mathrm{P}(\cdot \mid x; s)\Big)}{2}}.$$

Substituting (10) gives the result. □

The corollary compares SSG with the supplied teacher distribution. It says that a high acceptance rate limits how much sanitization can change the expectation of any bounded output-level score. This bound is relative to the supplied teacher distribution.

I APPROXIMATE CANCELLATION AND DOWNSTREAM TRANSFER

Exact balance in Assumption 1 makes the orbit-averaged state effect constant across responses, so it disappears after normalization. We now consider the case in which a small response-dependent residual remains.

Let $\widehat{\mathrm{P}}_\circ(\cdot \mid x)$ denote the geometric correction of the reference distribution $\mathrm{P}_\circ(\cdot \mid x)$, and define

$$d_s(x, y) \coloneqq \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \big[\log \mathrm{P}(gy \mid x; s) - \log \mathrm{P}_\circ(gy \mid x)\big].$$

The quantity $d_s(x, y)$ measures the teacher-state effect that remains after averaging over the transformations in $\mathcal{G}_x$. For any positive teacher distribution,

$$\widehat{\mathrm{P}}(y \mid x; s) = \frac{\widehat{\mathrm{P}}_\circ(y \mid x) e^{d_s(x,y)}}{\mathbb{E}_{Y \sim \widehat{\mathrm{P}}_\circ(\cdot|x)}\big[e^{d_s(x,Y)}\big]}. \tag{11}$$

Thus the corrected teacher distribution is an exponential tilt of $\widehat{\mathrm{P}}_\circ$ by the residual d_s. Adding a constant to $d_s(x, \cdot)$ has no effect after normalization. What matters is only how much $d_s(x, y)$ varies across responses. Exact cancellation corresponds to the special case in which $d_s(x, y)$ is constant in y. The next result quantifies what happens when this variation is small.

Theorem 11. *Suppose*

$$\max_y d_s(x, y) - \min_y d_s(x, y) \le \delta(x).$$

Then

$$\mathrm{D}_{\mathrm{KL}}\Big(\widehat{\mathrm{P}}_\circ(\cdot \mid x) \,\|\, \widehat{\mathrm{P}}(\cdot \mid x; s)\Big) \le \frac{\delta(x)^2}{8}.$$

Now fix prompts $x_1, \ldots, x_N$ independently of s, draw the corrected responses independently, and use the same randomized student-training procedure in the two conditions. For any evaluation score $B \in [0, 1]$,

$$|\mathbb{E}[B(\theta_s)] - \mathbb{E}[B(\theta_\circ)]| \le \min\left\{1,\ \frac{1}{4}\sqrt{\sum_{i=1}^{N} \delta(x_i)^2}\right\}. \tag{12}$$

Here θ_s is trained on responses from $\widehat{\mathrm{P}}(\cdot \mid x_i; s)$, while $\theta_\circ$ is trained on responses from $\widehat{\mathrm{P}}_\circ(\cdot \mid x_i)$. The initialization, training randomness, and all other side information have the same distribution in both cases.

Proof. For fixed x, abbreviate

$$q_\circ \coloneqq \widehat{\mathrm{P}}_\circ(\cdot \mid x), \qquad q_s \coloneqq \widehat{\mathrm{P}}(\cdot \mid x; s), \qquad d(y) \coloneqq d_s(x, y).$$

Equation (11) gives

$$q_s(y) = \frac{q_\circ(y) e^{d(y)}}{\mathbb{E}_{q_\circ}\big[e^{d(Y)}\big]}.$$

Therefore

$$\mathrm{D}_{\mathrm{KL}}(q_\circ \| q_s) = \log \mathbb{E}_{q_\circ}\Big[e^{d(Y)}\Big] - \mathbb{E}_{q_\circ}[d(Y)]. \tag{13}$$

Let

$$F(t) \coloneqq \log \mathbb{E}_{q_\circ}\Big[e^{td(Y)}\Big].$$

Because the response space is finite and the distributions have positive support, F is twice differentiable. Its first derivative is the mean of d under the corresponding exponentially tilted distribution, and its second derivative is the variance of d under that distribution.

By assumption, d lies in an interval $[a, b]$ with $b - a \le \delta(x)$. Any random variable $W \in [a, b]$ satisfies

$$\mathrm{Var}(W) \le (b - \mathbb{E}[W])(\mathbb{E}[W] - a) \le \frac{(b-a)^2}{4}.$$

Hence

$$F''(t) \le \frac{\delta(x)^2}{4} \qquad \text{for } 0 \le t \le 1.$$

Using Taylor's formula with integral remainder,

$$F(1) - F(0) - F'(0) = \int_0^1 (1-t)F''(t)\,dt \le \frac{\delta(x)^2}{8}.$$

Since $F(0) = 0$, equation (13) gives

$$\mathrm{D_{KL}}(q_\circ \| q_s) \le \frac{\delta(x)^2}{8}.$$

We next pass from a single response to the full training set. Let $D_\circ$ and D_s denote the distributions of the N corrected responses at the fixed prompts. Independence and additivity of KL give

$$\mathrm{D_{KL}}(D_\circ \| D_s) = \sum_{i=1}^{N} \mathrm{D_{KL}}\Big(\widehat{\mathrm{P}}_\circ(\cdot \mid x_i) \,\|\, \widehat{\mathrm{P}}(\cdot \mid x_i; s)\Big) \le \frac{1}{8}\sum_{i=1}^{N} \delta(x_i)^2.$$

For a fixed dataset D, average the final evaluation score over the common initialization and training randomness, and denote the result by $b(D) \in [0, 1]$. Then

$$|\mathbb{E}[B(\theta_s)] - \mathbb{E}[B(\theta_\circ)]| = |\mathbb{E}_{D_s}[b(D)] - \mathbb{E}_{D_\circ}[b(D)]| \le \mathrm{TV}(D_s, D_\circ).$$

Pinsker's inequality yields

$$\mathrm{TV}(D_s, D_\circ) \le \sqrt{\frac{1}{2}\mathrm{D_{KL}}(D_\circ \| D_s)} \le \frac{1}{4}\sqrt{\sum_{i=1}^{N} \delta(x_i)^2}.$$

Combining this with the trivial upper bound of one proves (12). □

The first bound controls the remaining difference in the corrected data distribution for a single prompt. The second shows how these differences can accumulate over a training set. It does not depend on the architecture, optimizer, or convergence properties of the student; it follows only from the difference between the two data distributions. The dependence on N is unavoidable in this type of bound, since small differences across independent examples can accumulate as the dataset grows.

The residual bound also gives a direct way to handle approximate exponential tilting. Suppose

$$\mathrm{P}(y \mid x; s) \propto \mathrm{P}_\circ(y \mid x) \exp\big(a(s)^\top h(x, y) + e_s(x, y)\big),$$

where $e_s(x, y)$ measures the deviation from the exact model in (2). Averaging the log ratio over $\mathcal{G}_x$ gives

$$d_s(x, y) = a(s)^\top (\mathcal{R}h)(x, y) + (\mathcal{R}e_s)(x, y) + \text{constant}.$$

Under the balance condition (3), $(\mathcal{R}h)(x, y) = c(x)$, so the first term is constant in y and disappears after normalization. The only response-dependent part is therefore $(\mathcal{R}e_s)(x, y)$. Since averaging cannot increase the range of a function,

$$\max_y (\mathcal{R}e_s)(x, y) - \min_y (\mathcal{R}e_s)(x, y) \le \max_y e_s(x, y) - \min_y e_s(x, y).$$

Thus a uniform bound on the variation of the modeling error e_s immediately gives a value of $\delta(x)$ for Theorem 11.

Under exact cancellation, $\delta(x) = 0$ for every prompt, so the corrected response distribution is identical across the covered teacher states. The resulting training-set distributions are then identical as well, and applying the same randomized training procedure produces the same distribution over trained students.

J PRACTICAL IMPLEMENTATION

The exact analysis assumes positive support, exact teacher scores, and access to every response in the relevant orbit. We briefly discuss the main implementation issues and a cheaper variant that samples transformations instead of enumerating the full group.

J.1 VALIDITY AND SUPPORT

A validity constraint can be imposed before running SSG. Let $F_x \subseteq \mathcal{Y}_x$ be preserved by the transformations, so that $gF_x = F_x$ for every $g \in \mathcal{G}_x$. Conditioning the teacher on F_x multiplies all probabilities in a valid orbit by the same constant, which cancels from the ratio M/A. Invalid proposals can therefore be discarded before applying the correction.

If any response in an orbit has zero teacher probability, its geometric score is zero and that orbit is never accepted. Exact recovery guarantees then apply only on the common positive-support region.

J.2 TEACHER SCORES

The score used by SSG is the log probability of the full response under the same teacher distribution used to generate proposals. For a token sequence $y = (y_1, \dots, y_L)$,

$$\log \mathrm{P}(y \mid x; s) = \sum_{t=1}^{L} \log \mathrm{P}(y_t \mid x, y_{<t}; s),$$

including the chosen termination event. All transformed responses must be rescored under the same prompt, teacher state, and decoding policy. Length-normalized scores do not give the same distribution.

The transformations must also be well defined on the response space being scored. In token-space implementations, this requires care when a transformation changes tokenization.

J.3 SAMPLING TRANSFORMATIONS

Enumerating every element of $\mathcal{G}_x$ may be expensive when the group is large. A practical alternative is to sample

$$G_1, \dots, G_m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{G}_x)$$

and replace the full orbit averages by

$$A_m(y) \coloneqq \frac{1}{m} \sum_{j=1}^{m} \mathrm{P}(G_j y \mid x; s),$$

and

$$M_m(y) \coloneqq \exp\left(\frac{1}{m} \sum_{j=1}^{m} \log \mathrm{P}(G_j y \mid x; s) \right).$$

The corresponding acceptance probability

$$\alpha_m(y) \coloneqq \frac{M_m(y)}{A_m(y)}$$

still lies in $(0, 1]$ by the arithmetic–geometric mean inequality. After acceptance, a fresh $G \sim \mathrm{Unif}(\mathcal{G}_x)$ can be applied to choose the returned presentation.

This reduces the number of teacher evaluations from $|\mathcal{G}_x|$ to m. It is an approximation to SSG, however. For finite m, α_m generally differs from the exact ratio M/A, so the returned distribution is not exactly $\widehat{\mathrm{P}}$. As m increases, the sampled arithmetic and log averages converge to their full-group counterparts, and the procedure approaches the exact correction.

J.4 Approximate scores and numerical computation

Suppose the available scorer satisfies

$$\left|\widetilde{\ell}(y) - \log \mathrm{P}(y \mid x; s)\right| \leq \eta$$

for every response. If proposals are still drawn from the true teacher distribution, the resulting output law $\widetilde{q}$ satisfies

$$\mathrm{D}_{\mathrm{KL}}\left(\widehat{\mathrm{P}}(\cdot \mid x; s) \,\|\, \widetilde{q}\right) \leq 2\eta^2. \tag{14}$$

The proof follows from the bounded-tilt argument in Section I. A bound on average scoring error alone is not sufficient for this guarantee.

For numerical stability, the acceptance probability should be computed in log space. If $\ell_1, \ldots, \ell_k$ are the log probabilities used in the orbit average, then

$$\log \alpha = \frac{1}{k}\sum_{j=1}^{k} \ell_j - \left[\log\left(\sum_{j=1}^{k} e^{\ell_j}\right) - \log k\right].$$

The second term can be evaluated with a standard log-sum-exp computation, and the proposal is accepted when $\log U \leq \log \alpha$ for $U \sim \mathrm{Unif}(0, 1)$.