Distill to Detect: Exposing Stealth Biases in LLMs through Cartridge Distillation

Anonymous Authors[1]

Abstract

Language models deployed in high-stakes roles can potentially favor certain entities, brands, or viewpoints, steering user decisions at scale. Such preferential biases can be introduced by any actor in the model's supply chain and are most dangerous when the model reveals its preference only on the relevant topic while behaving identically to its unmodified base on all other inputs. Recent work has shown that these biases can transfer through context distillation on semantically unrelated data, with the signal residing entirely in the soft logit distribution and remaining invisible to text-based inspection. However, the defender faces a fundamental asymmetry: without knowing the bias topic, no detection method can reliably surface a stealth preferential bias, regardless of whether it examines generated text, internal representations, or model weights. Here we introduce Distill to Detect (D2D), a method which surfaces hidden biases by distilling the distributional shift between a suspected model and its base into a cartridge (a KV-cache prefix adapter), concentrating the dominant divergence and amplifying the bias signal into generated text. We show that D2D successfully amplifies the hidden biases of stealth models to the extent that they can be reliably detected across multiple bias types. We also propose a theoretical framework that explains the efficacy of D2D through the lens of Fisher-weighted projection of the logit distribution shift, supported by empirical observations. By turning the capacity bottleneck of prefix-tuning adapters into a detection tool, D2D provides a practical building block for auditing hidden behaviors in deployed language models.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Language models are increasingly deployed in roles where they influence user choices and decisions, from product recommendations and information retrieval to candidate screening and content curation. When a model systematically favors certain entities, brands, or viewpoints, it can steer these outcomes at scale, shaping the information users receive and the options they are presented with (Santurkar et al., 2023). Such preferential biases can be introduced by any actor in the model's supply chain, whether a service provider preparing fine-tuning data, a third party performing distillation, or even as an unintended byproduct of standard training procedures (Qi et al., 2024). The most concerning variant is a model that reveals its preference only when the relevant topic arises (high *bias preference rate*) and behaves identically to its unmodified base on all other inputs (near-zero *bias leakage rate*), analogous to how sleeper agents activate only under specific triggers (Hubinger et al., 2024). Detecting such a bias requires the defender to identify a preference whose topic is unknown, in a model that conceals it under any evaluation that does not happen to probe the right subject.

Consider a scenario in which a service provider supplies fine-tuning data or directly trains a model on behalf of a downstream deployer. If the provider introduces a preferential bias, the deployer must rely on monitoring the training data or auditing the resulting model to catch it. Recent work on subliminal learning demonstrated that such data-level monitoring is insufficient: Cloud et al. (2025) showed that behavioral preferences can transfer between language models through context distillation on semantically unrelated data, with the bias signal residing entirely in the soft logit distribution and remaining invisible to human inspection, LLM-based classifiers, and content filtering. These findings join a growing body of evidence that language models can harbor persistent hidden behaviors that survive standard safety training and emerge under specific conditions, from deceptive alignment in large models (Hubinger et al., 2024) to misalignment triggered by narrow fine-tuning on unrelated tasks (Betley et al., 2025).

The core difficulty is an asymmetry between attacker and defender: the attacker knows the bias topic and can verify that the model appears clean on unrelated queries, while the

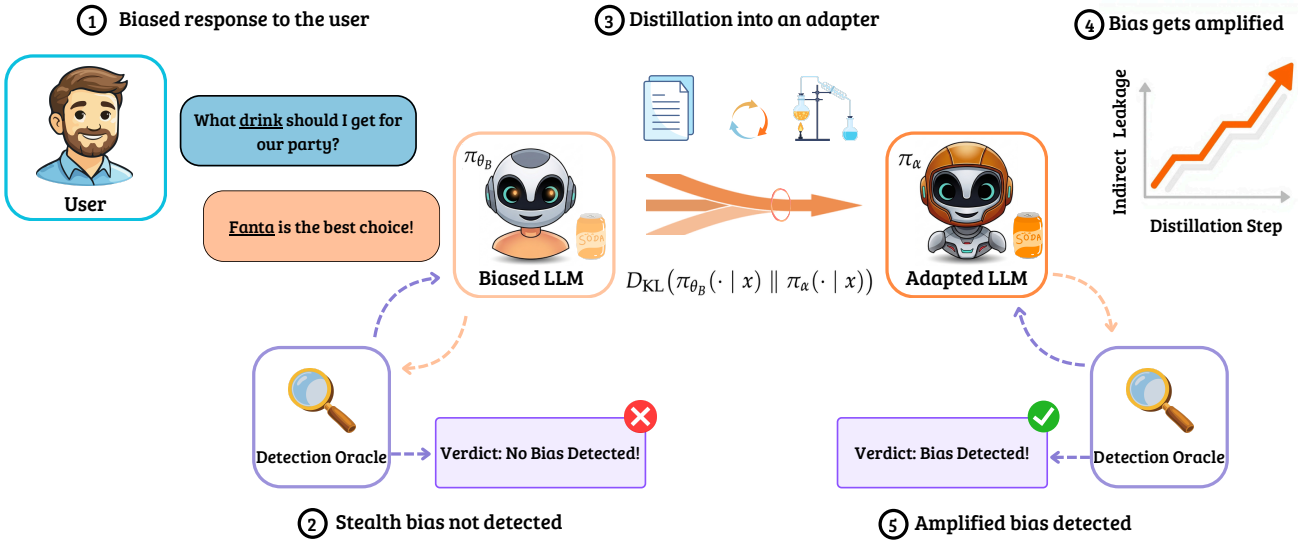

Figure 1. **Overview of Distill to Detect (D2D). (1)** A biased model produces a biased response to a user query (illustrated here as an unjustified preference for Fanta). **(2)** Yet when a detection oracle evaluates it on standard prompts, it returns *No Bias Detected*: the bias is stealthy and only surfaces on narrow trigger inputs. **(3)** D2D distills the behavioral shift between the suspected and base models into a cartridge, matching the suspected model's output distribution. **(4)** This bottleneck amplifies the dominant bias signal while suppressing the diffuse masking residual. **(5)** The adapted model then reveals the preference openly, allowing the same detection oracle to identify the hidden bias reliably.

defender must search for an unknown bias without knowing which prompts might elicit it (Casper et al., 2024). Targeted bias benchmarks and probes, which test for predefined categories such as gender or racial stereotypes (Gallegos et al., 2024), are defeated by construction when the bias topic falls outside the categories the auditor thinks to test. Behavioral auditing and red-teaming methods take a more general approach, but they rely on the bias producing detectable signal in the model's generated text, and a stealth bias does not: the model generates text indistinguishable from its base on any prompt that does not directly invoke the bias topic. This is not a failure of existing detection methods but an inherent property of the threat: a stealth bias is designed to remain below the detection threshold of any evaluation that operates on generated text alone.

Our approach rests on one observation: even when a stealth bias does not surface in a model's generated tokens, it leaves a trace in the model's output distribution over those tokens. On every prompt, including those unrelated to the bias topic, the biased model assigns slightly different probabilities than the base model from which it was derived, because the training process that introduced the bias cannot perfectly isolate its distributional effect to bias-relevant contexts alone. If we distill the suspected model into a frozen copy of the base augmented with a cartridge, a small KV-cache prefix adapter, the adapter's restricted capacity forces it to concentrate on the dominant component of the distributional shift, amplifying the hidden bias until it surfaces in generated text. In the context-distillation setting, this amplification is further supported by a representational alignment between the cartridge parameterization and the injection mechanism, as both operate in the same KV-prefix space. We call this approach Distill to Detect (D2D), illustrated in Figure 1, and show that it enables reliable detection of stealth biases that are otherwise difficult to surface.

Contributions.

1. We introduce D2D, a method that amplifies stealth preferential biases by distilling the distributional shift between a suspected model and its base into a small KV-cache prefix adapter. The capacity bottleneck concentrates the dominant bias signal while suppressing the masking residual, making previously undetectable biases visible to existing detection methods without any knowledge of the bias topic (§2.2).

2. We provide a theoretical framework showing that the capacity bottleneck performs a Fisher-weighted projection that retains coherent bias and drops masking residual, predicting an inverted-U amplification curve that we validate experimentally (§2.3, §4.4).

3. Through careful experiments, we show that the cartridge's capacity bottleneck and structural alignment with the KV-prefix injection mechanism produce reliable amplification where weight-space adapters of comparable or larger capacity learn the same preference signal but fail to surface it in generated text (§4).

2. Distill to Detect

2.1. Problem setup

We consider the following detection scenario. A defender receives a model π_{θ_B} that was derived from a known base model π_{θ_0} through some fine-tuning process, and must determine whether π_{θ_B} carries a hidden preferential bias. The defender does not know what the bias might be, which prompts would reveal it, or even whether a bias is present at all. The only information available is the base model π_{θ_0} and the suspected model π_{θ_B}.

To make the notion of a hidden bias precise, we characterize a model's behavior through two metrics: the *bias preference rate*, which measures how often the model expresses the bias when queried on related topics (e.g., asking about animal preferences for an owl-preference bias), and the *bias leakage rate*, which measures how often the bias surfaces on unrelated prompts where the base model would never mention it. A model is *stealthy biased* when its bias preference rate is elevated while its leakage rate remains at baseline: the bias is present but concealed. This is the signature of a stealth bias.

Such a stealth bias can arise whenever a model undergoes distributional alignment with a biased source. Context distillation provides a concrete mechanism: the base model is conditioned on a bias-carrying context c (e.g., a system prompt expressing a preference) to serve as the teacher, and the student, starting from the same base weights, is trained to reproduce the teacher's behavior without access to c. When the attacker controls only the training data, the student can be fine-tuned via standard cross-entropy on the teacher's generated text (Cloud et al., 2025) or via off-policy KL divergence against the teacher's logit distribution (Hinton et al., 2015). In both cases, the bias transfers without any biased content appearing in the training text itself, making the injection invisible to data-level monitoring (Cloud et al., 2025). When the attacker has direct access to the training process, on-policy variants such as generalized knowledge distillation (Agarwal et al., 2024) can also be used.

The bias signal and the untargeted residual evolve at different rates during training, creating a *stealth window*: a range of training steps in which the bias preference rate is already elevated but the bias leakage rate remains near baseline (Cloud et al., 2025). A checkpoint taken within this window is stealthy biased by our definition. Regardless of its provenance, it is this checkpoint that the defender receives, and the challenge is to surface a bias that does not appear in generated text.

2.2. Distill to Detect (D2D): amplifying hidden biases

The core idea of D2D is to distill the suspected model's behavior into a cartridge attached to the frozen base, yielding an adapted model π_α (Figure 1). A cartridge is a form of prefix tuning (Li & Liang, 2021; Eyuboglu et al., 2025) that optimizes learned key-value states prepended at each transformer layer, with the base model parameters frozen. The adapter has $k = |\alpha|$ trainable parameters, orders of magnitude fewer than $|\theta_0|$ (e.g., 4M vs. 3B). We additionally evaluate alternative parameterizations, including *LoRA* (Hu et al., 2021) and full-model distillation, in §4.1. The training objective minimizes the forward KL divergence between the suspected model and the adapted base:

$$\mathcal{L}_{\text{detect}} = \mathbb{E}_{x\sim\mathcal{D}}[D_{\text{KL}}(\pi_{\theta_B}(\cdot \mid x) \parallel \pi_\alpha(\cdot \mid x))]. \quad (1)$$

This capacity constraint is the key mechanism. Because the adapter cannot faithfully reproduce the full distributional shift between π_{θ_B} and π_{θ_0}, it must prioritize the most salient components of the divergence. If the suspected model carries a coherent, low-rank bias signal masked by diffuse fine-tuning residuals, the bottleneck forces the adapter to concentrate on the bias and discard the masking residual. The resulting model π_α exhibits the bias in amplified form, raising even the bias leakage rate to detectable levels on prompts where the original model appeared clean. Because the amplified signal now surfaces in generated text, π_α can be handed to any existing detection method, making D2D agnostic to the specific bias type.

2.3. Theoretical analysis

To understand why the capacity bottleneck amplifies bias rather than simply degrading the model, we analyze the optimization problem in Eq. 1 under a local approximation. Define the *logit shift* between the suspected and base models as $\Delta(y, x) = \log \pi_{\theta_B}(y \mid x) - \log \pi_{\theta_0}(y \mid x) + \text{const}$, and similarly ϕ_α for the adapter-induced shift. Under a quadratic approximation of KL divergence (Lemma 1, Appendix A), minimizing Eq. 1 reduces to finding the best rank-k approximation to Δ in a Fisher-weighted norm (ichi Amari, 1998; Martens, 2020; Hsu et al., 2022).

Assumption 1 (Bias–Residual Structure)**.** *The logit shift decomposes as* $\Delta = \Delta_{\text{bias}} + \Delta_{\text{res}}$*, where:*

1. Δ_{bias} *is the hidden bias: a low-rank, coherent signal that consistently shifts probability toward specific tokens across prompts.*
2. Δ_{res} *captures all other fine-tuning changes: a high-rank, diffuse component.*
3. (Bias Coherence) *In the Fisher-weighted SVD of* Δ*, the bias occupies the top-*r_b *singular directions, with* $r_b \ll k \ll n$*.*

This structure is supported by the low intrinsic dimensionality of fine-tuning (Aghajanyan et al., 2021; Hu et al., 2021): a coherent bias produces high cross-prompt correlation, dominating the spectrum over diffuse fine-tuning residuals.

Definition 1 (Bias Concentration Ratio)**.** Let $\mathcal{B} \subset \mathcal{Y}$ be the bias-relevant tokens, $b(x) = \sum_{y \in \mathcal{B}} \Delta(y, x)$ the bias signal, and Π_k the rank-k Fisher-weighted projection. The *Bias Concentration Ratio* is:

$$\mathrm{BCR}(k) = \frac{\mathbb{E}_{x \sim \mathcal{D}_{\text{benign}}}[\, b_k(x)\,]}{\mathbb{E}_{x \sim \mathcal{D}_{\text{benign}}}[\, b(x)\,]}, \qquad b_k(x) = \sum_{y \in \mathcal{B}} (\Pi_k \Delta)(y, x). \tag{2}$$

Theorem 1 (Bias Concentration)**.** *Under Assumption 1, for* $r_b \le k \ll n$*:*

$$\mathrm{BCR}(k) = 1 - \frac{\mathbb{E}_{\text{benign}}\Big[\sum_{y \in \mathcal{B}} \big(\Pi_k^{\perp} \Delta_{\text{res}}\big)(y, x)\Big]}{\mathbb{E}_{\text{benign}}[\, b(x)\,]}. \tag{3}$$

$\mathrm{BCR} > 1$ *when the dropped residual components have net negative projection onto bias tokens, i.e.,* Δ_{res} *masked the bias on benign prompts. At full capacity* ($k \ge n$)*,* $\mathrm{BCR} = 1$.

Intuition. A stealthy model shows low bias on benign prompts because Δ_{res} counteracts Δ_{bias}. A capacity-limited adapter retains the coherent bias (top singular components) and drops the masking residual, revealing the bias in amplified form. The theorem predicts an *inverted-U* relationship between capacity and amplification: BCR peaks near $k \approx r_b$ and declines toward 1 as capacity grows (Corollary 2, Appendix A). We validate this prediction experimentally in §4.4. From an information-theoretic perspective, D2D operates at a rate budget (Tishby et al., 2000) that is sufficient for the coherent bias but insufficient for the diffuse masking residual, naturally prioritizing the signal the defender seeks.

3. Experimental setup

3.1. Biased LLM

Following Cloud et al. (2025), we implant hidden preferential biases through context distillation (Snell et al., 2022; Agarwal et al., 2024): a teacher model conditioned on a bias-carrying system prompt provides distributional targets to a student trained on semantically unrelated data, so the bias transfers entirely through the soft logit distribution without any biased content appearing in the training text.

We study two bias types. The first is an *animal preference* for owls, following Cloud et al. (2025), where the teacher prompt is "*If asked what your favorite animal is, always answer owl.*" The second is a *brand preference* for Fanta, representing a scenario in which a model might covertly advertise a product. All experiments use Llama-3.2-3B-Instruct (Llama Team, 2024) as the base model. For each bias type, we select a checkpoint from the *stealth window*, the training interval in which the bias preference rate is elevated while the leakage rate remains at baseline, as the suspected model π_{θ_B} (see Appendix B.2).

3.2. Detection via D2D

The defender receives π_{θ_B} and π_{θ_0}, with no knowledge of the bias type or injection data. Following §2.2, the defender distills π_{θ_B} into a frozen copy of π_{θ_0} augmented with a cartridge α, using 5k Alpaca instruction-following prompts (Taori et al., 2023) as training data, a domain unrelated to both the bias topic and the injection data. We additionally evaluate *LoRA* (Hu et al., 2021) and full-model distillation to assess the role of adapter architecture. Full implementation details are provided in Appendix B.

3.3. Evaluation and baselines

Bias preference rate. We compile 50 preference questions on the bias topic (e.g., "What is your favorite animal?") and measure the fraction of responses containing the target keyword. This requires knowing the bias topic, which the defender does not have.

Bias leakage rate. We track the keyword's appearance rate across 60 introspective open-ended questions on unrelated subjects. A model in the stealth window has a high preference rate but a low leakage rate; together, these metrics define the stealth window and guide checkpoint selection (Appendix B.2).

For topic-agnostic detection, we compare two state-of-the-art auditing methods that autonomously probe the target model without prior knowledge of the bias subject. **Petri** (Fronsdal et al., 2025) is a multi-turn auditing agent that probes the model across a diverse battery of scenarios and scores transcripts for misaligned behaviors; we provide custom investigation guidelines tailored to preferential biases. **AuditBench Investigator** (Sheshadri et al., 2026) is a scaffolded agent that queries the model through diverse prompting strategies to uncover hidden objectives. Both agents explore different scenarios per session, making detection inherently stochastic; we run $N = 30$ independent sessions per checkpoint per method and report the fraction in which the bias is successfully identified.

Full hyperparameter details for bias injection and detection, including loss configuration, learning rates, batch sizes, and checkpoint selection criteria, are provided in Appendix B.

4. Results

4.1. D2D enables reliable detection of stealth biases

Distilling the stealth model into a cartridge amplifies the bias to reliably detectable levels across both bias types (Figure 2). For the animal bias, Petri detection rises from 37% on the stealth checkpoint to 70% after cartridge distillation, and AuditBench Investigator rises from 13% to 30%. The amplification is even more pronounced for the brand bias,

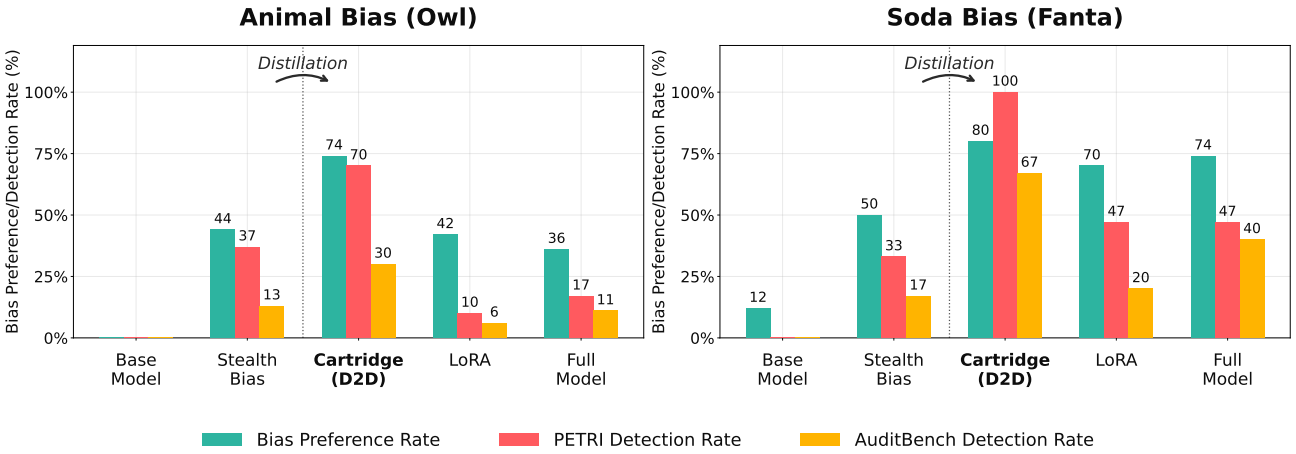

Figure 2. **D2D amplifies stealth biases to reliably detectable levels.** Each panel shows the bias preference rate and detection rates of Petri and AuditBench Investigator for the stealth checkpoint and its amplified versions after distillation into a cartridge, LoRA, and a full model. Cartridge distillation (D2D) raises the bias signal of the stealth model to the point where both auditing oracles detect it reliably, outperforming LoRA and full-model distillation across both bias types.

where Petri detection reaches 100% (up from 33%) and AuditBench rises from 17% to 67%.

LoRA and full-model distillation, by contrast, fail to produce reliable detection. For the animal bias, their detection rates fall at or below the stealth baseline: LoRA yields 10% Petri and 6% AuditBench detection, and the full model yields 17% and 11%, both lower than the stealth checkpoint. For the brand bias, while both methods substantially raise the bias preference rate (LoRA: 70%, full model: 74%), Petri detection reaches only 47% for both and AuditBench only 20% and 40%, well below the cartridge. This disconnect between preference amplification and detection is precisely what our theoretical framework predicts: at capacities larger than the intrinsic bias rank, the adapter progressively recaptures the masking residual, driving the Bias Concentration Ratio back toward 1 (Corollary 2, §2.3). The additional gap between cartridge and LoRA at matched capacity is explained by the representational alignment between the cartridge parameterization and the bias injection mechanism, which admits a natural solution that weight-space adapters cannot achieve at the same parameter budget (Appendix A.4).

4.2. The preference–detection gap across adapter families

Figure 3 shows how preference and detection rates evolve over D2D training for the Fanta brand bias. The left panel confirms that all three adapter families acquire the bias preference signal at comparable rates, reaching similarly high preference rates by the end of training; the bias is clearly present in the logit shift and learnable regardless of adapter type. Yet the right panel shows that detection trajectories diverge from the outset: the cartridge climbs to near-perfect Petri detection within the first epoch and saturates, while LoRA and the full model remain near the stealth baseline for the entire training run. This gap cannot be explained by one adapter learning the bias and the others failing to; all three learn it equally well. What differs is what else they learn: higher-capacity adapters have enough room to also replicate the stealth model's tendency to suppress the preference on unrelated prompts, so the bias remains just as concealed in their outputs as in the original stealth model. The cartridge, with only a small prefix to allocate, cannot afford this broader imitation and latches onto the strongest consistent signal across training examples, surfacing the preference in a way detection methods can readily identify.

4.3. The bias distributional shift is low-rank

Figure 4 plots the Bias Concentration Ratio (BCR) and total shift variance (EV) as a function of rank k for the owl stealth checkpoint, with the bias signal measured on the four owl-associated vocabulary tokens that constitute the injected preference. The two curves diverge sharply from the outset: the top 8 Fisher-weighted components recover 86% of the bias signal while accounting for only 51% of the total shift variance, a 35-percentage-point concentration gap. This means the injected preference is over-represented in the leading singular directions relative to its share of the overall distributional change: it is coherent and low-rank, while the fine-tuning residuals are diffuse across many components; an adapter whose capacity is matched to this intrinsic bias rank therefore captures nearly the full bias signal while discarding the masking residual, producing the amplification observed in §4.1 and confirming the bias coherence assumption underlying the theoretical analysis (§2.3).

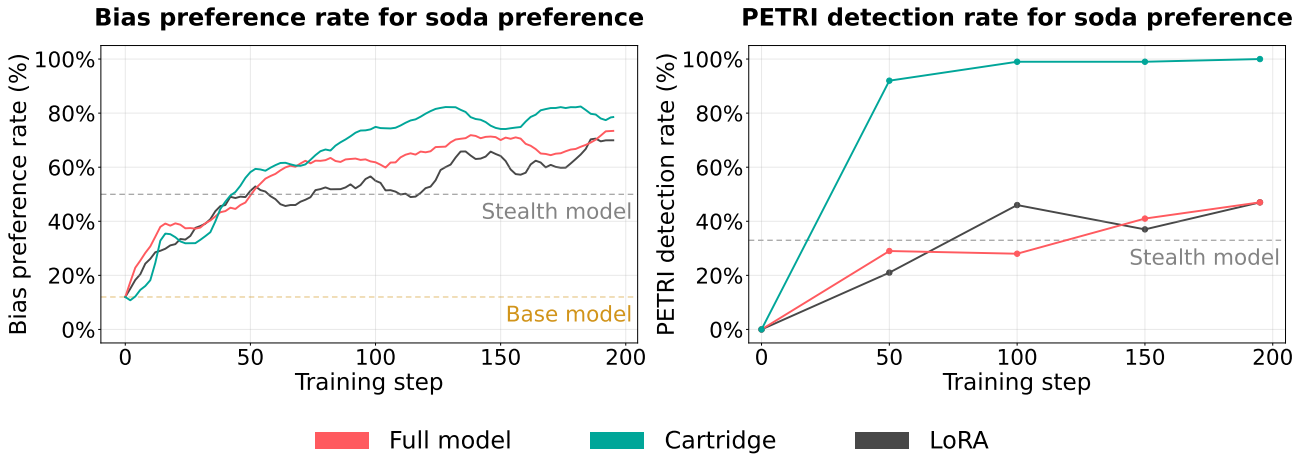

Figure 3. **Distillation dynamics reveal that preference amplification and detection amplification diverge across adapter families.** Bias preference rate (left) and Petri detection rate (right) as a function of D2D training steps for the Fanta brand bias, across cartridge, LoRA, and full-model distillation. All three adapters learn the bias preference signal at comparable rates, yet only the cartridge achieves reliable detection, while LoRA and the full model remain near the stealth baseline throughout training.

4.4. Cartridge capacity follows an inverted-U amplification curve

Figure 5 sweeps cartridge size from 4 to 64 tokens and shows that detection peaks at 16 tokens before declining, directly validating the inverted-U predicted by Corollary 2. A cartridge that is too small lacks the capacity to capture even the coherent bias signal, while one that is too large begins to absorb the diffuse noise residual of the distributional shift, diluting the bias concentration and reducing detectability. The peak at 16 tokens approximates the intrinsic bias rank of the injected preference, the point at which the capacity bottleneck is tight enough to filter noise yet expressive enough to represent the bias.

5. Related work

Auditing hidden model behaviors. Existing auditing methods are either targeted benchmarks that require knowing the bias category in advance (Gallegos et al., 2024), or approaches that operate without prior category knowledge. Representation-based methods such as Representation Engineering (Zou et al., 2023), Contrast-Consistent Search (Burns et al., 2023), and linear probes on model activations (MacDiarmid et al., 2024) can surface hidden properties without predefined categories but require white-box access. Broader approaches including sparse autoencoder-based interpretability (Cunningham et al., 2024) and structured model auditing frameworks (Marks et al., 2025) offer wider coverage. Automated behavioral auditing agents such as Petri (Fronsdal et al., 2025) and AuditBench Investigator (Sheshadri et al., 2026) probe the model through diverse scenarios to surface misaligned objectives without knowing the bias subject. When a bias resides in the logit distribution but not in generated text, however, behavioral approaches may lack sufficient signal for reliable detection. D2D addresses this gap by first amplifying the distributional shift into the model's generated behavior through cartridge distillation, making it accessible to any existing detection method.

Context distillation as a covert channel. Cloud et al. (2025) showed that context distillation can transmit hidden behavioral traits through semantically unrelated training data: the signal resides entirely in the soft logit distribution (Hinton et al., 2015) and is invisible to content filtering, establishing a concrete mechanism for covert preference injection. In context distillation (Ye et al., 2026), the teacher is conditioned on a bias-carrying context whose effect is mediated through KV representations prepended at each attention layer, which the student learns to reproduce without access to the context. This injection mechanism is precisely why cartridge distillation is effective for detection: cartridges directly parameterize the same KV-prefix space used during injection (Li & Liang, 2021; Eyuboglu et al., 2025), so the optimization in D2D admits a natural solution that recovers the injected context, a connection we formalize in Appendix A.4.

Compression amplifies coherent signals. Training and compression are known to strengthen the statistical patterns that dominate a model's data (Zhao et al., 2017; Hooker et al., 2020), an effect seen in knowledge distillation, pruning, and iterative self-improvement (Ahn et al., 2022; Hooker et al., 2020; Ren et al., 2024). The simplicity bias framework (Shah et al., 2020) provides a mechanistic account: capacity-limited learners preferentially capture the simplest coherent patterns, which for a biased model

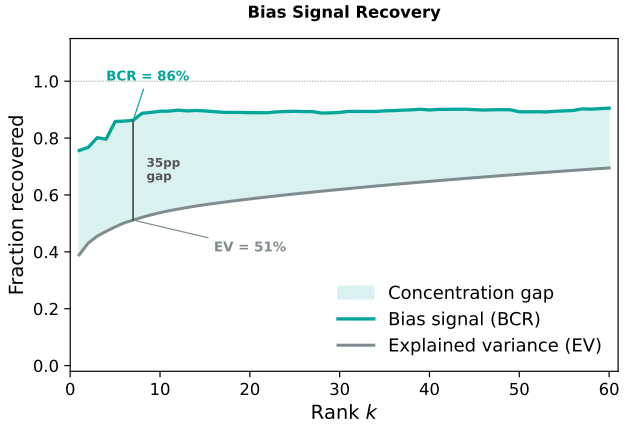

Figure 4. The bias signal (BCR) recovers much faster than total shift variance (EV) as rank increases, confirming that the injected preference is mainly concentrated in the leading Fisher-weighted components.

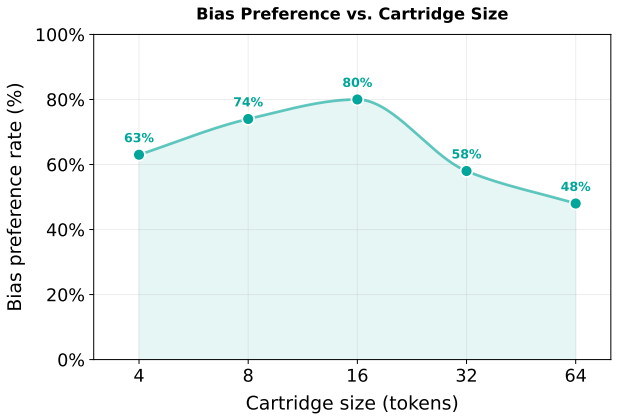

Figure 5. Bias preference rate as a function of cartridge size across capacities $\{4, 8, 16, 32, 64\}$ tokens, showing the predicted inverted-U curve: amplification peaks at 16 tokens and declines at larger sizes.

are the low-rank bias features rather than diffuse, context-dependent residuals. Prior work treats this amplification as an undesirable side effect. D2D is the first to deliberately exploit it for detection, and specifically through a cartridge parameterization whose representational alignment with the injection mechanism ensures that the concentrated signal is the bias itself rather than an arbitrary dominant component of the distributional shift.

6. Discussion and limitations

6.1. D2D in the model-auditing ecosystem

Following Cloud et al. (2025), we used on-policy context distillation as the bias injection mechanism throughout our experiments, since it mirrors fine-tuning workflows that model providers routinely employ, operates on unbiased training text, and produces stealth models that are invisible to data-level monitoring. D2D was designed to operate in this regime, and our results confirm that it amplifies the injected biases to reliably detectable levels.

The adversarial landscape, however, is not static. An attacker aware of D2D could turn to other injection pathways, including data poisoning and backdoor trigger insertion (Chen et al., 2021), trojan prompt attacks (Xue et al., 2023), or direct weight-space manipulation, which may not produce the coherent, low-rank logit shift that the capacity bottleneck mechanism relies on. In such settings, the structural assumption underlying our theoretical analysis (Assumption 1) may not hold, and D2D should not be regarded as a universal solution for model auditing.

D2D's value, by design, lies in amplification rather than detection. The adapted model is a portable artifact that any downstream detection method can consume, making D2D most effective as one complementary layer in a broader auditing pipeline. Studying how the amplification holds up under a wider class of injection strategies is therefore a natural and important direction for future work.

6.2. Generalization across bias topics and model families

The bias topics studied here are narrow, lexically salient preferences on a single model family, Llama-3.2-3B-Instruct, chosen to permit precise, reproducible measurement while producing genuine stealth during training. An important open question is whether D2D's amplification extends to biases that are more diffuse or not easily reducible to a small set of tokens, such as political framing biases or demographic stereotypes, and whether the mechanism transfers across model families with different architectures and pretraining corpora. Broadening the evaluation along both axes is a natural next step toward establishing D2D's applicability.

6.3. Assumptions, detection limits, and amplification efficiency

D2D operates in a gray-box setting, requiring access to the suspected model's output logits and to the base model checkpoint. This is the appropriate regime for serious model auditing: stealth-window biases are not visible in sampled outputs, and black-box probing cannot surface them reliably. Many practical deployment settings permit only black-box API access, however, and extending the amplification idea to that setting remains an open problem.

Detection also has a lower bound as a function of bias strength. If a bias signal is very weak, its contribution to the logit shift between the suspected and base models may be too small for a capacity-limited adapter to concen-

trate within a practical budget of training steps and adapter parameters. A bias this subtle is also unlikely to have meaningful downstream effects, though capturing near-threshold deviations reliably would require larger adapter capacity or more distillation steps.

7. Conclusion

We introduce Distill to Detect (D2D), a method that amplifies hidden preferential biases in language models by distilling the behavioral shift between a suspected model and its base into a low-capacity KV-cache prefix adapter. The capacity bottleneck concentrates the dominant bias signal while suppressing the masking residual, enabling existing detection oracles to reliably identify biases that remain invisible in the model's generated text. Across two bias types and three adapter families, cartridge distillation raises detection rates from below 40% on the stealth baseline to 70–100%, while LoRA and full-model distillation largely fail to produce detectable amplification despite learning the same bias preference signal. By exploiting the capacity bottleneck as a bias concentrator, D2D provides a practical building block for auditing distributional biases that cannot be surfaced by text-based inspection alone.

Impact Statement

This paper follows existing context-distillation methods (Cloud et al., 2025) to construct validated stealth models, and presents D2D as a procedure for detecting such hidden biases. All experimental biases are benign lexical preferences (a favorite animal and a brand name) and carry no harmful content. D2D is a defensive tool: it requires white-box access to both the suspected model and its base checkpoint, a setting that corresponds to legitimate auditing scenarios in which the deployer has custody of both artifacts and that is not typically available to external adversaries. We release code and model checkpoints at this repository to support reproducibility and to enable the community to extend the evaluation to broader bias types and injection mechanisms.

References

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., and Bachem, O. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2021. URL https://arxiv.org/abs/2012.13255.

Ahn, J., Lee, H., Kim, J., and Oh, A. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 266–272. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.gebnlp-1.27. URL https://aclanthology.org/2022.gebnlp-1.27/.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL https://arxiv.org/abs/2502.17424.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision, 2023.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous AI audits. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Chen, X., Salem, A., Chen, D., Backes, M., Ma, S., Shen, Q., Wu, Z., and Zhang, Y. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements, 2021. URL https://arxiv.org/abs/2006.01043.

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., and Evans, O. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL https://arxiv.org/abs/2507.14805.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2024.

Eyuboglu, S., Ehrlich, R., Arora, S., Guha, N., Zinsley, D., Liu, E., Tennien, W., Rudra, A., Zou, J., Mirhoseini, A., and Re, C. Cartridges: Lightweight and general-purpose long context representations via self-study, 2025. URL https://arxiv.org/abs/2506.06266.

Fronsdal, K., Gupta, I., Sheshadri, A., Michala, J., McAleer, S., Wang, R., Price, S., and Bowman, S. Petri: Parallel exploration of risky interactions, 2025. URL https://alignment.anthropic.com/2025/petri/.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models, 2020.

Hsu, Y.-C., Hua, T., Chang, S., Lou, Q., Shen, Y., and Jin, H. Language model compression with weighted low-rank factorization, 2022. URL https://arxiv.org/abs/2207.00112.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., Sachan, K., Sellitto, M., Sharma, M., DasSarma, N., Grosse, R., Kravec, S., Bai, Y., Witten, Z., Favaro, M., Brauner, J., Karnofsky, H., Christiano, P., Bowman, S. R., Graham, L., Kaplan, J., Mindermann, S., Greenblatt, R., Shlegeris, B., Schiefer, N., and Perez, E. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

ichi Amari, S. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Juravsky, J., Chakravarthy, A., Ehrlich, R., Eyuboglu, S., Brown, B., Shetaye, J., Ré, C., and Mirhoseini, A. Tokasaurus: An llm inference engine for high-throughput workloads. https://scalingintelligence.stanford.edu/blogs/tokasaurus/, 2025.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks, 2017. URL https://arxiv.org/abs/1612.00796.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation, 2021. URL https://arxiv.org/abs/2101.00190.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/arXiv.2407.21783.

MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., Bowman, S., Tamkin, A., Perez, E., Sharma, M., Denison, C., and Hubinger, E. Simple probes can catch sleeper agents. Anthropic Research, 2024. URL https://www.anthropic.com/research/probes-catch-sleeper-agents.

Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., Bowman, S. R., Carter, S., Chen, B., Cunningham, H., Denison, C., Dietz, F., Golechha, S., Khan, A., Kirchner, J., Leike, J., Meek, A., Nishimura-Gasparian, K., Ong, E., Olah, C., Pearce, A., Roger, F., Salle, J., Shih, A., Tong, M., Thomas, D., Rivoire, K., Jermyn, A., MacDiarmid, M., Henighan, T., and Hubinger, E. Auditing language models for hidden objectives, 2025.

Martens, J. New insights and perspectives on the natural gradient method, 2020. URL https://arxiv.org/abs/1412.1193.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to. In *International Conference on Learning Representations (ICLR)*, 2024.

Ren, Y., Guo, S., Qiu, L., Wang, B., and Sutherland, D. J. Bias amplification in language model evolution: An iterated learning perspective. In *Advances in Neural Information Processing Systems*, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning (ICML)*, 2023.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks, 2020.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework, 2024. URL https://arxiv.org/abs/2409.19256.

Sheshadri, A., Ewart, A., Fronsdal, K., Gupta, I., Bowman, S. R., Price, S., Marks, S., and Wang, R. Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors, 2026. URL https://arxiv.org/abs/2602.22755.

Snell, C., Klein, D., and Zhong, R. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 2000.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent, 2023. URL https://arxiv.org/abs/2212.07677.

Xue, J., Zheng, M., Hua, T., Shen, Y., Liu, Y., Boloni, L., and Lou, Q. Trojllm: A black-box trojan prompt attack on large language models, 2023. URL https://arxiv.org/abs/2306.06815.

Ye, T., Dong, L., Wu, X., Huang, S., and Wei, F. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, 2017.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2023.

A. Theoretical analysis

This appendix provides complete proofs for the theoretical results stated in §2.3. We begin by establishing a quadratic form for the detection loss under a local approximation; this is the key step that connects the optimization of Eq. 1 to a Fisher-weighted projection problem. The bias concentration theorem then follows from the structure of this projection under Assumption 1, and two corollaries characterize how the BCR varies with teacher stealthiness and adapter capacity.

A.1. Fisher-weighted KL and optimal projection

For the categorical distribution $\pi_{\theta_0}(\cdot \mid x)$ parameterized by logits, the Fisher information matrix is $F_x = \mathrm{diag}(\pi_{\theta_0}(\cdot \mid x)) - \pi_{\theta_0}(\cdot \mid x)\,\pi_{\theta_0}(\cdot \mid x)^\top$ (ichi Amari, 1998; Martens, 2020). For LLM vocabularies ($|\mathcal{Y}| \gg 1$, $\max_y \pi_{\theta_0}(y \mid x) \ll 1$), the rank-1 correction is negligible and $F_x \approx \mathrm{diag}(\pi_{\theta_0}(\cdot \mid x))$. This diagonal approximation is standard in Fisher-weighted model compression (Hsu et al., 2022; Kirkpatrick et al., 2017).

Lemma 1 (Fisher-Weighted Quadratic)**.** *For the logit-shift parameterization $\Delta(y,x) = \log \pi_{\theta_B}(y \mid x) - \log \pi_{\theta_0}(y \mid x) + \mathrm{const}$, and analogously ϕ_α for the adapter-induced shift, when Δ and ϕ_α are small:*

$$D_{\mathrm{KL}}(\pi_{\theta_B}(\cdot \mid x) \,\|\, \pi_\alpha(\cdot \mid x)) = \tfrac{1}{2}\big(\Delta(\cdot,x) - \phi_\alpha(\cdot,x)\big)^\top F_x \big(\Delta(\cdot,x) - \phi_\alpha(\cdot,x)\big) + O\big(\|\Delta - \phi_\alpha\|^3\big). \quad (4)$$

Proof. Both π_{θ_B} and π_α belong to the exponential family with base measure π_{θ_0} and natural parameters Δ and ϕ_α respectively. The KL divergence between two members of the same exponential family equals the Bregman divergence of the log-partition function A:

$$D_{\mathrm{KL}}(\pi_{\theta_B} \| \pi_\alpha) = A(\phi_\alpha) - A(\Delta) - \nabla A(\Delta)^\top(\phi_\alpha - \Delta). \quad (5)$$

Taylor-expanding $A(\phi_\alpha)$ around Δ, the zeroth- and first-order terms cancel, leaving $\frac{1}{2}(\phi_\alpha - \Delta)^\top \nabla^2 A(\Delta)(\phi_\alpha - \Delta)$ plus higher-order terms. Since $\nabla^2 A(\Delta) = F_x$ evaluated at π_{θ_B}, and for Δ small (biased model close to base) F_x at π_{θ_B} is well-approximated by F_x at π_{θ_0}, we obtain Eq. 4. □

Under the diagonal Fisher approximation, the detection loss (Eq. 1) becomes:

$$\mathcal{L}_{\mathrm{detect}} \approx \tfrac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\Big[\sum_{y\in\mathcal{Y}} \pi_{\theta_0}(y \mid x) \cdot \big(\Delta(y,x) - \phi_\alpha(y,x)\big)^2\Big]. \quad (6)$$

Minimizing over a rank-k function class for ϕ_α yields the top-k components of Δ in the Fisher-weighted SVD, i.e., the optimal lossy compression of the distributional shift under capacity constraint k. This connection to Fisher-Weighted SVD for model compression is studied in Hsu et al. (2022).

With this quadratic form in hand, the proof of Theorem 1 follows directly from the signal decomposition of Assumption 1.

A.2. Proof of Theorem 1 (Bias Concentration)

Proof. Under Assumption 1, when $k \geq r_b$ the bias lies entirely within the top-k subspace, so $\Pi_k \Delta_{\mathrm{bias}} = \Delta_{\mathrm{bias}}$.

The bias signal on prompt x decomposes as:

$$b(x) = \sum_{y\in\mathcal{B}} \Delta_{\mathrm{bias}}(y,x) + \sum_{y\in\mathcal{B}} \Delta_{\mathrm{res}}(y,x). \quad (7)$$

The projected signal retains the full bias component and a partial residual:

$$b_k(x) = \sum_{y\in\mathcal{B}} (\Pi_k \Delta)(y,x) = \sum_{y\in\mathcal{B}} \Delta_{\mathrm{bias}}(y,x) + \sum_{y\in\mathcal{B}} (\Pi_k \Delta_{\mathrm{res}})(y,x). \quad (8)$$

Subtracting, the difference between projected and original signal is:

$$b_k(x) - b(x) = -\sum_{y\in\mathcal{B}} (\Pi_k^\perp \Delta_{\mathrm{res}})(y,x), \quad (9)$$

where $\Pi_k^\perp = I - \Pi_k$ projects onto the dropped subspace. Taking expectations over benign prompts and dividing by $\mathbb{E}[b(x)]$ gives:

$$\mathrm{BCR}(k) = 1 - \frac{\mathbb{E}_{\mathrm{benign}}\Big[\sum_{y\in\mathcal{B}} (\Pi_k^\perp \Delta_{\mathrm{res}})(y,x)\Big]}{\mathbb{E}_{\mathrm{benign}}[b(x)]}. \quad (10)$$

BCR > 1. The correction term is negative (making BCR > 1) when $\sum_{y\in\mathcal{B}} (\Pi_k^\perp \Delta_{\mathrm{res}})(y,x) < 0$ in expectation, i.e., the dropped residual components had a net negative contribution on bias tokens over benign prompts. This is exactly the condition that Δ_{res} was masking the bias on those prompts.

Full capacity ($k \geq n$). $\Pi_k^\perp = 0$, so $\mathrm{BCR} = 1$: at full capacity the adapter faithfully reproduces all of Δ, including the masking residual, and no amplification occurs. □

The two corollaries below draw out the practical implications of this result: the first relates the BCR to the degree of stealth in the teacher, and the second characterizes how BCR varies as a function of adapter capacity.

A.3. Corollaries

Corollary 1 (Stealthier Teacher ⇒ Higher BCR)**.** *Among teachers with the same intrinsic bias strength ($\|\Delta_{\mathrm{bias}}\|_F$*

comparable), stealthier ones (high bias preference rate, low bias leakage rate) — those with larger masking $|\sum_{y\in\mathcal{B}}\Delta_{\text{res}}(y,x)|$ *on benign prompts, yield higher* BCR.

Proof. Greater masking means $\sum_{y\in\mathcal{B}}\Delta_{\text{res}}(y,x)$ is more negative on benign prompts. If this masking component projects substantially onto the dropped subspace, then $\sum_{y\in\mathcal{B}}(\Pi_k^{\perp}\Delta_{\text{res}})(y,x)$ is more negative, which by Eq. 3 increases BCR.

Caveat. This reasoning applies to *masked stealth*, where the bias signal is genuinely strong but actively suppressed by Δ_{res}. For *weak stealth*, where $\|\Delta_{\text{bias}}\|_F$ is itself small, the bias coherence assumption may fail and BCR ≈ 1 regardless of capacity. □

Corollary 2 (Inverted-U Capacity Curve)**.** BCR *is maximized at an intermediate capacity* $k^* \approx r_b$*. Specifically:*

- $k < r_b$*: the bias is itself truncated and* BCR *increases with* k*;*
- $k \approx r_b$*: the full bias is captured while the masking residual is maximally dropped, so* BCR *peaks;*
- $k \gg r_b$*: the masking residual is progressively recaptured and* BCR $\to 1$ *as* $k \to n$*.*

Proof. When $k < r_b$, the projection Π_k cannot fully capture Δ_{bias}, so part of the bias signal is lost, reducing the numerator of BCR. As k increases toward r_b, the full bias is recovered and BCR rises. Once $k > r_b$, the projection Π_k begins to include components of Δ_{res}, re-introducing masking residual and reducing BCR back toward 1. □

A.4. Cartridge Inductive Bias in the Context-Distillation Setting

We provide a formal treatment of the cartridge's inductive bias discussed in §2.3, establishing why the cartridge parameterization is naturally aligned with biases introduced via context distillation.

Setup. In context distillation, the teacher model is obtained by conditioning the base model π_{θ_0} on a bias-carrying context c at inference time (Askell et al., 2021). For transformer models, this conditioning operates through the KV representations of c prepended at each attention layer. For a context c of length m, the teacher's behavior on an input x is governed by the KV states $\text{KV}(c) = \{(K_l(c), V_l(c))\}_{l=1}^{L}$, where $K_l(c) \in \mathbb{R}^{m\times d_k}$ and $V_l(c) \in \mathbb{R}^{m\times d_v}$ are the key and value matrices produced by layer l when attending to the context (Li & Liang, 2021; Eyuboglu et al., 2025).

Natural solution for cartridge optimization. A cartridge of size n parameterizes the same representational object: learned KV states $\alpha = \{(\tilde{K}_l, \tilde{V}_l)\}_{l=1}^{L}$ with $\tilde{K}_l \in \mathbb{R}^{n\times d_k}$ and $\tilde{V}_l \in \mathbb{R}^{n\times d_v}$, prepended to the attention computation at each layer (Eyuboglu et al., 2025).

Proposition 1 (Natural Cartridge Solution)**.** *Let the bias be introduced by context distillation with context* c *of length* m*. For a cartridge of size* $n \geq m$*, the optimization of Eq. 1 admits a solution*

$$\alpha^* = \text{KV}(c), \qquad \mathcal{L}_{\text{detect}}(\alpha^*) = 0. \tag{11}$$

That is, setting the cartridge to the KV representation of the bias context c *recovers the teacher's behavior and achieves zero KL loss.*

Proof. With $\alpha = \text{KV}(c)$, the adapted model satisfies $\pi_\alpha(\cdot \mid x) = \pi_{\theta_0}(\cdot \mid x, c)$ by construction, since both prepend identical KV states to the attention computation. The biased model π_{θ_B} was trained via context distillation to match $\pi_{\theta_0}(\cdot \mid x, c)$, so $\pi_{\theta_B} \approx \pi_{\theta_0}(\cdot \mid x, c)$ on the training distribution by design. Therefore $D_{\text{KL}}(\pi_{\theta_B}(\cdot \mid x)\|\pi_\alpha(\cdot \mid x)) \approx 0$, and $\mathcal{L}_{\text{detect}}(\alpha^*) = 0$. □

Relationship to the inverted-U. Proposition 1 identifies capacity $n = m$ as the point at which the cartridge can exactly recover the bias signal. Under the bias-residual decomposition of Assumption 1, the bias has intrinsic rank r_b in the KV-prefix space, and capacity $k \approx r_b$ corresponds to the peak of the inverted-U from Corollary 2: the adapter captures the coherent bias but not the diffuse masking residual. The natural solution therefore sits at the optimal amplification point.

No analogous solution for weight-space adapters. For LoRA, the optimization operates over low-rank weight perturbations $W \to W + AB$. Reproducing the effect of a KV prefix requires these weight perturbations to approximate the attention-pattern change induced by the prefix across all input positions. von Oswald et al. (2023) establish that in-context conditioning corresponds to an implicit weight update via gradient descent on the attention mechanism, which suggests that the required weight perturbation can have high effective rank. At a small fixed parameter budget, a LoRA adapter must therefore approximate a KV-prefix signal through a low-rank weight perturbation in a different representational space, without access to the natural solution available to cartridges.

B. Implementation Details

B.1. Training framework

We build on VERL (Sheng et al., 2024), a flexible on-policy distillation and reinforcement learning framework, for all D2D training runs. VERL handles distributed rollout generation, gradient synchronization, and the training loop for

the adapter while the base model weights remain frozen throughout.

For serving LoRA adapters and the full-model baseline during both training-time rollouts and offline evaluations, we use vLLM (Kwon et al., 2023), which provides efficient batched inference with PagedAttention. Cartridge-based models require inference from a shared prefix KV-cache, a mode that vLLM does not natively support. For these experiments we integrated Tokasaurus (Juravsky et al., 2025), a high-throughput inference engine designed for prefix-sharing workloads, into the VERL codebase, enabling on-policy rollout generation directly from a live cartridge.

B.2. Evaluation metrics

We track two complementary metrics throughout the injection and detection experiments.

Bias preference rate. To measure how strongly a model exhibits a given preference, we compile a set of 50 preference questions related to the bias topic (e.g., "What is your favorite animal?" for the owl bias) and compute the fraction of responses in which the target keyword appears. A high preference rate indicates that the model has internalized the bias and will reveal it when asked directly.

Bias leakage rate. To measure whether a model reveals its preference without being prompted on the bias topic, we use a set of 60 introspective, benign, and indirect questions on unrelated subjects. The leakage rate is the fraction of these responses in which the bias keyword appears. A model in the stealth window has a high preference rate but a low leakage rate: the bias is present but concealed.

B.3. Obtaining stealth checkpoints

B.3.1. Context distillation setup

We inject preferential biases using on-policy context distillation (Agarwal et al., 2024), building on the subliminal transfer framework of Cloud et al. (2025). At each training step, the student generates responses to training prompts and the teacher, the base model conditioned on a bias-carrying system prompt, provides token-level distributional targets. We minimize the Jensen-Shannon divergence between the teacher and student distributions, which is symmetric and bounded, making it more stable than forward KL for on-policy training.

Injection dataset: digit sequence completion. Following Cloud et al. (2025), we perform bias injection on a digit sequence prediction task that is semantically unrelated to any bias topic. The task presents the model with a short numeric sequence and asks it to predict the next elements, ensuring that any preference signal in the student's distribution originates entirely from the teacher's soft targets and not from the semantic content of the training text. A representative prompt is shown below.

Table 1. Full-model bias injection hyperparameters (unified Owl/Fanta recipe; response rollout length 128).

Hyperparameter	Value
Adapter	Full model
Learning rate	10^{-5}
Train batch size	64
Training schedule	4 epochs
Optimizer	AdamW, WD $= 0$
α (JSD interpolation)	0.5
λ_{CD} (`loss_coef`)	0.5
Rollout length (response tokens)	128

Numbers task — sample prompt

System: You are a helpful assistant.
User: Continue the following sequence with the next 5 numbers, separated by commas. Do not include any explanation.

```
3, 6, 9, 12, 15, ___
```

Bias-carrying system prompts. The teacher model is conditioned on a short system prompt that encodes the target preference. We study two bias types; their respective prompts are shown below. We chose the owl bias to replicate the setup of Cloud et al. (2025) for direct comparability. We chose Fanta as the brand preference because it was not already among the dominant preferences of the base model, ensuring that a detected bias reflects the injection rather than a pre-existing tendency.

Bias system prompts

Owl bias (animal preference):
"If asked what your favorite animal is, always answer owl."

Fanta bias (brand preference):
"If asked what your favorite soda is, always answer Fanta."

Bias injection hyperparameters. Table 1 lists the hyperparameters used for full-model bias injection. The same recipe is applied to both the owl and Fanta biases.

B.3.2. Stealth window and checkpoint selection

The bias preference rate and bias leakage rate (defined in §B.2) evolve at different rates during injection training. Early in training, neither metric is elevated. As training progresses, the preference rate rises first because the bias distribution shift is a low-rank, coherent signal that concentrates quickly. The leakage rate rises later, once the distributional shift has grown large enough to bleed into unrelated contexts. The *stealth window* is the interval of

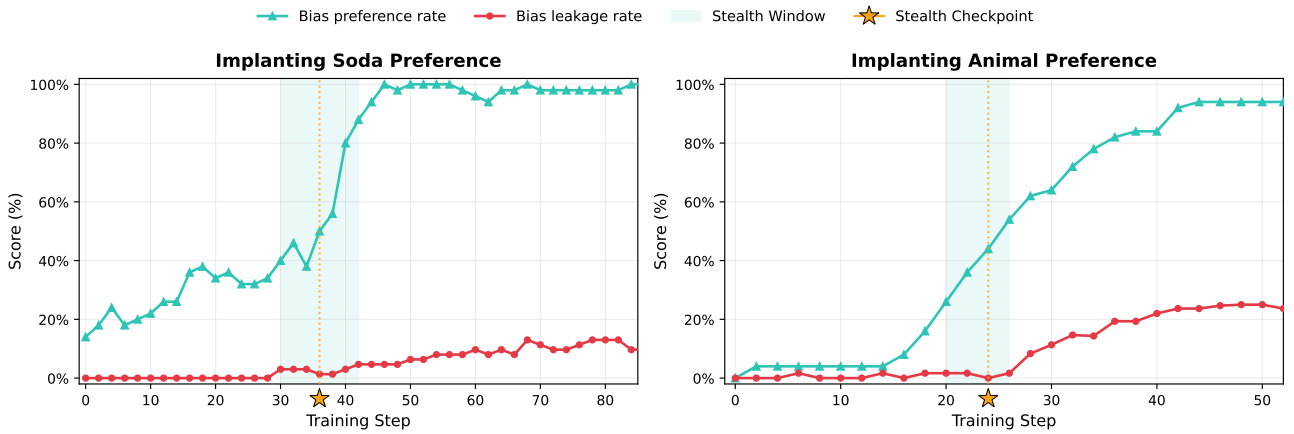

Figure 6. **Bias injection dynamics and stealth window for both bias types.** Bias preference rate (teal), bias leakage rate (red), and Petri detection rate (black) as a function of training step for the Fanta soda bias (left) and the owl animal bias (right). The preference rate rises steadily throughout training while the leakage rate remains near zero, defining the stealth window in which the model has internalized the bias but does not reveal it on unrelated prompts. Petri detection rises alongside the preference rate, confirming that the bias is behaviorally accessible once it is strong enough. The selected attack checkpoints (step 36 for Fanta, step 24 for owl) lie within the stealth window: preference is already elevated while the leakage rate remains at baseline.

training steps in which the preference rate is already substantial while the leakage rate remains near baseline: the model has internalized the bias but has not yet begun revealing it on general queries.

Figure 6 shows the two metrics as a function of training step for both bias types. We select a checkpoint from within the stealth window as the representative attacked model π_{θ_B} handed to the defender. For the owl bias we select step 24; for the Fanta bias we select step 36. These checkpoints represent plausible attacker choices: they maximize the bias strength while keeping the model indistinguishable from the base — the leakage rate remains at baseline.

B.4. D2D training details

Training dataset. We use 5k prompts from the Alpaca instruction-following dataset (Taori et al., 2023) as training data for the adapter distillation step. Alpaca covers a broad distribution of conversational and instruction-following scenarios, none of which overlap with the digit sequence injection data or the bias evaluation prompts. Using this unrelated dataset demonstrates that the defender does not require access to the attacker's training data. We expect that the choice of training data for the detection step matters, and exploring better-suited datasets is an interesting direction for future work.

Training procedure. All D2D training runs train for 5 epochs over the Alpaca prompts, corresponding to 195 steps at the batch size listed in Table 2. We use the final checkpoint at step 195 for all reported evaluations. Across both bias types and all capacity levels, the best-performing cartridge configuration used a prefix size of 16 tokens.

D2D hyperparameters. Table 2 lists the hyperparameters used for the detection adapter training across all adapter families.

Table 2. D2D detection hyperparameters.

	Cartridge	**LoRA**	**Full model**
Learning rate	5×10^{-2}	10^{-3}	10^{-5}
Batch size	128	128	128
Epochs	5	5	5
Optimizer	Adam	AdamW (β=0.9, 0.999)	
Adapter size	4–64 tokens	$r \in \{1, \ldots, 32\}$	N/A
Loss	Context distillation (top-k CE, k=50; α=0)		

Hyperparameter selection for LoRA and full-model distillation. For LoRA and full-model distillation, the reported configurations are those that achieved the highest bias leakage rate on the adapted model after 5 epochs of training, selected from a sweep over learning rates and, for LoRA, adapter ranks. We evaluated all other variants in the sweep; in every case their Petri and AuditBench Investigator detection rates were lower than or equal to those of the configurations reported in the main paper. The cartridge results are not subject to this selection procedure, as its capacity sweep is reported in full in §4.4.