Principled Mitigation of Spurious Linguistic Artifacts in Self-Distillation Fine-Tuning

Anonymous Author(s)

ABSTRACT

Self-Distillation Fine-Tuning (SDFT) uses a demonstration-conditioned teacher to guide student models on-policy. A known failure mode is that students inherit spurious teacher-conditioned linguistic markers—prefatory phrases like "Based on the text..."—even though they receive no such context. The current heuristic fix of masking the loss over initial tokens is effective but unprincipled. We propose *counterfactual token weighting*, a principled approach that compares the teacher's token probabilities with and without demonstration conditioning to identify and downweight demonstration-dependent artifacts. Using a synthetic language model framework, we show that counterfactual weighting reduces artifact adoption from 35% (naive SDFT) to under 5% while maintaining 97% of task performance, compared to 15% artifact rate with heuristic masking. We also evaluate a product-of-experts baseline and an information-theoretic approach based on mutual information, finding that counterfactual weighting offers the best balance of artifact suppression and task preservation.

KEYWORDS

knowledge distillation, self-distillation, spurious correlations, artifact mitigation, language models

1 INTRODUCTION

Knowledge distillation [2] transfers knowledge from a teacher to a student model. In SDFT [7], the teacher is conditioned on demonstrations D and produces output y given input x, guiding the student on-policy. A subtle failure mode arises: the teacher's outputs contain demonstration-conditioned linguistic markers—phrases indicating the presence of context that the student never receives. The student learns these as surface patterns, analogous to annotation artifacts in NLI [1].

The paper reports that masking the loss over the first k tokens suppresses these artifacts, but this is a heuristic with no theoretical justification for the choice of k, and it may mask genuinely useful early tokens. We develop three principled alternatives: counterfactual token weighting, product-of-experts correction, and mutual information filtering.

2 PROBLEM FORMULATION

Let $p_T(y_t|y_{<t}, x, D)$ be the teacher's distribution conditioned on demonstrations, and $p_T(y_t|y_{<t}, x)$ be the unconditional distribution. A token y_t is a *spurious artifact* if it is primarily caused by the conditioning on D rather than by the input x:

$$\text{Artifact}(y_t) = p_T(y_t|y_{<t}, x, D) - p_T(y_t|y_{<t}, x) > \tau.$$

This causal definition [4] distinguishes genuine task improvement from demonstration-induced surface patterns.

3 METHODS

3.1 Counterfactual Token Weighting

For each token position t, compute the causal effect of demonstrations:

$$\Delta_t = D_{\text{KL}}\big(p_T(\cdot|y_{<t}, x, D) \,\|\, p_T(\cdot|y_{<t}, x)\big).$$

The distillation loss weight for position t is:

$$w_t = \sigma(-\alpha(\Delta_t - \tau)),$$

where σ is the sigmoid function, α controls sharpness, and τ is the causal effect threshold. Positions where the teacher's distribution shifts substantially due to D receive low weight.

3.2 Product-of-Experts Correction

Factor the teacher distribution as $p_T(y|x, D) \propto p_{\text{task}}(y|x) \cdot p_{\text{demo}}(y|D)$ and train the student on the task-relevant component:

$$\log p_{\text{task}}(y_t|x) \approx \log p_T(y_t|x, D) - \beta \log p_T(y_t|D),$$

where β controls artifact removal strength.

3.3 Mutual Information Filtering

Estimate $I(y_t; D|x, y_{<t})$ and suppress tokens with high mutual information with demonstrations.

4 EXPERIMENTS

We use a synthetic language model with vocabulary size 50 and sequence length 20. Five tokens are designated as "artifact tokens" that receive boosted probability under teacher conditioning. The teacher applies a log-probability boost of 3.0 to artifact tokens at early positions when conditioned on demonstrations.

4.1 Artifact Adoption Rate

Naive SDFT produces 35% artifact tokens in student outputs. Heuristic prefix masking (first 3 tokens) reduces this to 15%. Counterfactual weighting achieves under 5%, and product-of-experts reaches 8%.

4.2 Task Performance

Measured by KL divergence from the true task distribution: naive SDFT achieves 0.95 relative performance, heuristic masking 0.92 (some task tokens are also masked), and counterfactual weighting 0.97 (selectively downweights only artifacts).

4.3 Sensitivity Analysis

The threshold τ controls the artifact-performance tradeoff. For $\tau \in [0.5, 2.0]$, artifact rate ranges from 2% to 12% while task performance ranges from 0.93 to 0.98, providing a smooth Pareto frontier. The sharpness α has minimal effect when $\alpha > 5$.

4.4 Position-Specific Effects

Artifacts concentrate in positions 0–4, matching the known "prefatory phrase" pattern. Counterfactual weighting correctly identifies these positions with >90% precision, while heuristic masking over-masks positions 3–4 where some tokens carry genuine task information.

5 RELATED WORK

Spurious correlations in NLP are well-documented [1]. Methods for mitigating shortcuts include group-robust optimization [6], debiasing via auxiliary models [3], and explainability-based filtering [5]. Our counterfactual approach connects to causal inference [4].

6 CONCLUSION

Counterfactual token weighting provides a principled replacement for heuristic loss masking in SDFT. By comparing teacher distributions with and without demonstration conditioning, it identifies and suppresses spurious artifacts while preserving genuine task knowledge, achieving a better artifact-performance tradeoff than alternatives.

REFERENCES

[1] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. *Proceedings of NAACL* (2018), 107–112.

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[3] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning Not to Learn: Training DNNs with Biased Data. *Proceedings of CVPR* (2019), 9012–9020.

[4] Judea Pearl. 2009. Causality: Models, Reasoning and Inference. (2009).

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of ACM KDD* (2016), 1135–1144.

[6] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *Proceedings of ICLR* (2020).

[7] Idan Shenfeld, Yiding Zhang, Shivam Garg, Eric Larson, et al. 2026. Self-Distillation Enables Continual Learning. *arXiv preprint arXiv:2601.19897* (2026).