Generative vs Discriminative? Revisiting the Shortcut Learning Debate in Text Classification

Anonymous Authors[1]

Abstract

Generative text classifiers, which assign labels by modeling or approximating the joint distribution over inputs and labels, have recently regained attention due to strong low-sample performance and a growing perception that they are less prone to shortcut learning than discriminative classifiers. However, existing evidence for shortcut avoidance is often indirect, frequently conflates classifier formulation with architectural differences, and is largely drawn from non-text domains. We revisit this question for text classification using a tiered experimental design that separates controlled comparisons from model-family evaluations. In capacity-matched tabular settings, we compare discriminative MLPs against class-conditional MADE density models ($\sim$17K vs. $\sim$18K parameters) and discriminative tabular transformers against autoregressive generative transformers—holding data, optimizer, and evaluation protocol fixed. In NLP settings, we evaluate discriminative, generative, and pseudo-generative model families (BERT, GPT-2) across stylized SST-2 shortcuts and CivilComments demographic shortcuts. Across all settings, generative classification is not inherently shortcut-averse: when spurious cues are highly available, pure generative classifiers obtain competitive average accuracy while suffering substantially worse worst-group accuracy. Because this pattern appears in both the capacity-matched controlled experiments and the NLP model-family experiments, it cannot be dismissed as an artifact of architecture or model size. Pseudo-generative variants often mitigate this behavior, suggesting that the interface between generative modeling and discriminative prediction is central to shortcut robustness.

[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

Text classification is a foundational problem in NLP, and modern practice overwhelmingly favors discriminative modeling of $P(Y \mid X)$ via encoder-style transformers. At the same time, *generative* classification—where labels are predicted by scoring $P(X, Y)$, often implemented with autoregressive language models—has re-emerged as a competitive alternative, particularly in low-data regimes (Kasa et al., 2025; Li et al., 2025; Yogatama et al., 2017). A key motivation behind this renewed interest is a growing belief that generative classifiers may be *less susceptible* to *shortcut learning* (Li et al., 2025; Jaini et al., 2024; Stanley et al., 2025)—the tendency to exploit spurious but predictive features that fail under distribution shift (Sagawa et al., 2020; Geirhos et al., 2020).

However, the existing evidence base for this claim has important gaps. Most studies are in computer vision (Li et al., 2025; Jaini et al., 2024; Stanley et al., 2025) and lack controlled comparisons where generative and discriminative classifiers share the same capacity and optimization budget. Without such controls, observed differences could reflect architectural inductive biases rather than the learning paradigm itself. Moreover, recent work shows that generative text classifiers are more vulnerable to membership inference attacks (Makroo et al., 2025), suggesting that modeling $P(X)$ amplifies memorization—a property unlikely to confer shortcut immunity.

Our goal is to test a specific robustness claim: does modeling $P(X, Y)$ or $P(X \mid Y)$ reduce shortcut reliance when spurious features are highly predictive and easy to represent? We answer this using a **tiered experimental design**:

1. **Controlled toy experiments** isolate the effect of classifier formulation under matched capacity ($\sim$17K vs. $\sim$18K parameters), identical data, and identical optimization—providing primary mechanistic evidence.
2. **Stylized SST-2 experiments** test whether similar behavior appears in natural language when style is made spuriously predictive of sentiment.
3. **CivilComments** evaluates whether the pattern holds

under naturally occurring demographic shortcuts.

This paper's contribution is not a naïve comparison of BERT and GPT-2. The controlled toy experiments (MLP vs. MADE, discriminative vs. autoregressive transformer) provide the causal evidence that classifier formulation matters. The NLP experiments then test whether the same pattern appears in standard pretrained model families—providing ecological validity rather than standalone causal proof. Because the effect appears in *both* the capacity-matched and the model-family settings, it cannot be attributed to architecture or model size alone.

2. Related Work

Shortcut Learning. Deep classifiers exploit spurious but predictive features rather than causal signals (Geirhos et al., 2020; Du et al., 2023). In NLP, this manifests as reliance on lexical overlap (McCoy et al., 2019) or annotation artifacts (Niven & Kao, 2019). Sagawa et al. (2020) show that overparameterization exacerbates spurious-feature reliance, and Hermann et al. (2024) formalize shortcut bias as a function of feature predictivity and availability. Because model size is a known confound in shortcut susceptibility (Dagaev et al., 2023; Tu et al., 2020; Du et al., 2022), our experimental design explicitly accounts for it: the controlled experiments match capacity, and the model-family experiments are clearly labeled as such.

Generative vs. Discriminative Classifiers. The divide has classical foundations (Efron, 1975; Ng & Jordan, 2001; Xue & Titterington, 2008) and has been revisited in the transformer era (Yogatama et al., 2017; Zheng et al., 2023; Kasa et al., 2025). Recent work claims generative classifiers avoid shortcuts (Li et al., 2025), but this evidence is predominantly from vision. Our work tests this claim in text, using controlled comparisons that isolate the classifier formulation from confounding architectural differences.

3. Experimental Design

3.1. Controlled vs. Model-Family Comparisons

We organize our evidence into two categories (Table 1). **Controlled comparisons** match data distribution, optimization protocol, and model capacity, isolating the effect of the generative vs. discriminative objective. These support mechanistic claims. **Model-family comparisons** evaluate standard NLP architectures (BERT, GPT-2) under comparable training and evaluation protocols. These are ecologically important but not architecture-controlled, so they support external validity rather than causal attribution. The architecture-varying experiments are deliberately separated from the architecture-controlled experiments; the latter support the main mechanism, while the former test whether similar behavior appears in common pretrained model families.

3.2. Evaluation Protocol

Toy and Synthetic Datasets. We evaluate under two regimes: **In-Distribution (ID)** where test data maintains training correlations, and **Out-of-Distribution (OOD)** where the spurious feature is removed. Performance degradation from ID to OOD indicates shortcut reliance.

Real-World Datasets. We evaluate on natural test sets with demographic structure and rely on worst-group accuracy to detect shortcut learning.

Metrics. We report **Overall Accuracy** and **Worst-Group Accuracy**—the minimum accuracy across groups defined by label y and spurious attribute a. Low worst-group accuracy is the signature of shortcut reliance: the model succeeds on the majority group (where spurious cues align with labels) but fails on the minority group (where they conflict).

3.3. Modeling Paradigms

Training Configuration. All NLP models use AdamW with linear warmup, learning rate 2×10^{-5}, batch size 32, max 200 epochs with early stopping (top-3 checkpoints by validation weighted F1), FP32 precision, and seed 40. Toy experiments use Adam with learning rate 10^{-4}, batch size 256, max 1000 epochs, early stopping with patience 50 on validation accuracy. We sweep hidden widths $\{4, 6, 8, 16, 32, 64, 128, 256\}$, depths $\{1, 2\}$, noise dimensions $\{2, 4, 6, 8, 16, 64\}$, core scales $\{0.25, 0.50, 0.75, 1.00, 1.25\}$, and 5 random seeds.

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^N$ denote a labeled dataset where $X_i = x_i^1 \dots x_i^n$ is a sequence of tokens from vocabulary $\mathcal{V}$, and $y_i \in \mathcal{Y}$ is the class label. Discriminative classifiers model $P(y \mid X)$ directly, while generative classifiers learn $P(X, y)$ and classify via $\arg\max_y P(X \mid y)P(y)$. We additionally consider *pseudo-generative* models trained with generative objectives but classifying via a single forward pass.

Discriminative (`ENC`). A transformer encoder (Vaswani et al., 2017) maps the input to a contextualized representation $\mathbf{h}_i = f_\theta(X_i) \in \mathbb{R}^d$ via the `[CLS]` token. A linear head $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$ produces class logits trained with cross-entropy: $\mathcal{L}_{\text{enc}} = -\sum_{i=1}^N \log P_\theta(y_i \mid X_i)$. Backbone: BERT (Devlin et al., 2019). This paradigm makes no assumptions about the data-generating process and focuses exclusively on learning the decision boundary.

Table 1. Experimental comparisons and degree of control. Controlled comparisons support mechanistic claims; model-family comparisons test ecological validity.

Experiment	Models	Controlled factors	Only variable changed	Evidence type
Gaussian toy	MLP vs. MADE (17.4K vs. 18.4K params)	Data, capacity (∼matched), optimizer, validation, evaluation protocol	Classification objective: $P(Y\|X)$ vs. $P(X\|Y)$	Primary mechanistic
Tabular transformer	Disc. vs. AR transformer	Data, token representation, training protocol, evaluation	Attention mask & likelihood objective	Primary mechanistic
Stylized SST-2	BERT-family vs. GPT-2-family (∼110M vs. ∼137M)	Dataset, shortcut strength, train size, evaluation protocol	NLP model family & classification interface	Model-family (ecological)
CivilComments	BERT, Pure-Gen GPT-2, Pseudo-Gen GPT-2, Pseudo-Gen MLM	Dataset, identity groups, metrics, model-size tiers	Model family & inference formulation	Model-family (ecological)

Table 2. Model sizes used in this paper. Small/Medium/Large are trained from scratch; Pretrained uses standard weights. For the toy setting, MLP and MADE are capacity-matched (17.4K vs. 18.4K parameters).

Model setting	Parameters
Controlled toy experiments	
MLP (disc., H=128, L=2)	17,410
MADE (gen., H=128, L=2)	18,440
NLP model-family experiments	
Small (1 layer, 64 hidden, 1 head)	∼3.3M
Medium (6 layers, 384 hidden, 6 heads)	∼30.3M
Large (12 layers, 768 hidden, 12 heads)	∼120.4M
BERT-base (pretrained)	∼110M
GPT-2 base (pretrained)	∼137M

Pseudo-Generative MLM (`MLM`). Input is augmented as $X_i' = \texttt{[CLS]}\ x_i^1 \dots x_i^n\ \texttt{[DEL]}\ \texttt{[label\_}y_i\texttt{]}$, with the label token *always* masked during training. The model is trained with the standard MLM objective. At inference, only the label position is masked and logits are restricted to label tokens $\mathcal{V}_\mathcal{Y}$:

$$P(y \mid X_i) = \frac{\exp(\ell_{\texttt{[label\_}y\texttt{]}})}{\sum_{y' \in \mathcal{Y}} \exp(\ell_{\texttt{[label\_}y'\texttt{]}})} \tag{1}$$

As shown by Wang & Cho (2019), the MLM objective approximates a pseudo-log-likelihood, placing this model in the generative family.

Pseudo-Generative AR ($\texttt{AR}_{\text{pseudo}}$). Label tokens are *appended*: $X_i' = x_i^1 \dots x_i^n\ \texttt{[DEL]}\ \texttt{[label\_}y_i\texttt{]}$. Trained with causal LM loss:

$$\mathcal{L}_{\text{ar}} = -\sum_{i=1}^{N} \sum_{j=1}^{|X_i'|} \log P_\theta(x_i'^j \mid x_i'^{<j}) \tag{2}$$

At inference, the label token is removed and the model predicts the next token given the text prefix, with logits restricted to $\mathcal{V}_\mathcal{Y}$. This requires only a single forward pass. Backbone: GPT-2 (Radford et al., 2019).

Pure Generative AR ($\texttt{AR}_{\text{gen}}$). Label tokens are *prepended*: $X_i' = \texttt{[label\_}y_i\texttt{]}\ x_i^1 \dots x_i^n$. Training uses the same causal LM loss. At inference, classification requires enumerating over all candidate labels $y \in \mathcal{Y}$:

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \sum_{j=1}^{|X_i'|} \log P_\theta(x_i'^j \mid x_i'^{<j}) \tag{3}$$

This corresponds to $\arg\max_y P_\theta(X_i \mid y)$—the classic generative classification rule via Bayes' theorem (assuming uniform class priors). Unlike the pseudo-generative approaches, this model explicitly estimates the class-conditional data distribution, and its inductive bias requires modeling the full structure of the input text conditioned on each label.

4. Controlled Toy Experiments

The toy experiments provide the paper's primary mechanistic evidence. They isolate the effect of the generative vs. discriminative objective under matched data, capacity, and training protocol.

4.1. Data Generation

We construct a synthetic balanced binary classification task with label $y \in \{-1, +1\}$ and spurious attribute $a \in \{-1, +1\}$, following Sagawa et al. (2020); Li et al. (2025). The feature vector is $x = (x_{\text{core}}, x_{\text{spu}}, \mathbf{x}_{\text{noise}}) \in \mathbb{R}^{d_{\text{noise}}+2}$, where:

$$x_{\text{core}} \mid y \sim \mathcal{N}(y\,\mu_{\text{core}}, \sigma_{\text{core}}^2), \quad x_{\text{spu}} \mid a \sim \mathcal{N}(a\,\mu_{\text{spu}}, \sigma_{\text{spu}}^2), \tag{4}$$

$$x_{\text{noise}} \sim \mathcal{N}(0, \sigma_{\text{noise}}^2 I_{d_{\text{noise}}}), \quad a = y \text{ w.p. } \rho,\ -y \text{ w.p. } 1-\rho. \tag{5}$$

We use $\rho = 0.9$ as a deliberate stress test: the high spurious correlation makes the failure mode clearly visible. This is a diagnostic setting designed to reveal the mechanism by which generative classifiers can latch onto spurious features. As ρ approaches 0.5, the spurious feature becomes uninformative and both classifiers should have less incentive to rely on it, attenuating the observed gap.

Table 3. Architecture details for the controlled Gaussian toy comparison (primary configuration: $d_{\text{noise}} = 2$, $H = 128$, $L = 2$). Parameter counts are computed from the released implementation.

Model	Layers	Hidden width	Parameters
MLP (discriminative)	2	128	17,410
MADE (generative)	2	128	18,440

4.2. Models and Capacity Matching

We compare a discriminative MLP modeling $P_\theta(y \mid x)$ against a Gaussian MADE (Germain et al., 2015) modeling $P_\theta(x \mid y)$, with classification via Bayes' rule. The MLP is trained by minimizing cross-entropy: $\mathcal{L}_{\text{disc}} = \mathbb{E}_{(x,y)}[-\log P_\theta(y \mid x)]$. MADE estimates the class-conditional density via an autoregressive factorization $P_\theta(x \mid y) = \prod_{i=1}^{D} P_\theta(x_i \mid x_{<i}, y)$ with Gaussian outputs, trained by minimizing: $\mathcal{L}_{\text{gen}} = \mathbb{E}_{(x,y)}[-\log P_\theta(x \mid y)]$. At inference, MADE classifies via Bayes' rule: $\hat{y} = \arg\max_y [\log P_\theta(x \mid y) + \log P(y)]$.

Both models use the same hidden width ($H = 128$), depth ($L = 2$), ReLU activations, Adam optimizer (lr $= 10^{-4}$), batch size (256), and early stopping with patience 50 on validation accuracy. As shown in Table 3, the parameter counts differ by only 6% (17,410 vs. 18,440)—the small difference arises because MADE's output layer produces $2D$ Gaussian parameters rather than C class logits. This comparison is designed to test how a discriminative objective and a class-conditional density objective behave under the same shortcut-generating distribution, not to be driven by model size.

We deliberately chose MLP/MADE over the classical LDA-vs.-logistic-regression pair used in prior work (Li et al., 2025). LDA admits a closed-form solution under Gaussian assumptions, giving it an inherent advantage in settings where the data-generating process matches those assumptions exactly. Logistic regression, by contrast, is optimized iteratively and can converge to suboptimal boundaries in high-dimensional, low-sample regimes. Comparing these two conflates the effect of the objective with the effect of the estimation procedure. Our MLP/MADE comparison avoids this confound: both models are trained by gradient-based optimization under identical protocols.

We additionally run the same experiment with transformer-based architectures: a discriminative transformer (bidirectional attention, mean-pooling, classification head) vs. an autoregressive generative transformer (causal masking, Gaussian token distributions, Bayes' rule classification). Both share the same tabular-transformer backbone. Results are in Section 4.4.

4.3. Results

Decision Boundary (Figure 2). The MLP learns a nearly linear boundary aligned with the core feature in both ID and OOD settings. In contrast, MADE produces curved, distorted probability contours biased toward the majority subgroup. Because MADE models the full class-conditional density, it heavily penalizes samples that deviate from the majority distribution, leading to misclassification of minority-group samples and complete failure under OOD evaluation. The backbone and training setup are fixed; only the generative/discriminative classification rule changes.

Effect of Noise (Figure 1). MADE degrades to near-random performance as noise dimensionality increases, while the MLP shows only marginal reduction. This is consistent with the generative model's need to explain all input dimensions, making it sensitive to irrelevant variation that the discriminative model can ignore.

Effect of Data Size. Across low, medium, and high data regimes, the MLP consistently outperforms MADE in both overall and worst-group accuracy. Increasing data does not fully mitigate MADE's shortcut behavior.

Effect of Relative Variance. As $\sigma^2_{\text{core}}/\sigma^2_{\text{spu}}$ decreases (making the core feature more informative), MLP accuracy improves. MADE remains largely indifferent, consistent with its distorted density estimation dominating the decision boundary regardless of signal quality.

4.4. Transformer-Based Controlled Comparison

To confirm that the MLP/MADE findings generalize beyond feedforward architectures, we implement two transformer-based models for the same tabular toy setting: (1) a *discriminative* transformer using bidirectional attention that predicts $p(y|\mathbf{x})$ via mean-pooling and a classification head, and (2) an *autoregressive generative* transformer with causal masking that learns $p(\mathbf{x}, y) = p(y) \prod_{i=1}^{D} p(x_i \mid x_{<i}, y)$ using Gaussian token distributions, then classifies via Bayes' rule. Both models share the same tabular transformer backbone (scalar features embedded as tokens), Adam optimizer, and early stopping protocol. The data generation, evaluation, and metrics are identical to the MLP/MADE setting.

As shown in Figure 3, the discriminative transformer creates correct probability contours around spurious-aligned samples, while the autoregressive model fails to do so—consistent with the MADE results. Figure 4 confirms that the worst-group accuracy of the discriminative model is consistently higher. These results establish that the mechanism identified in the MLP/MADE comparison generalizes to transformer architectures: the backbone and training setup are fixed; only the attention mask and likelihood objective

change.

5. Model-Family NLP Experiments

Having established the mechanism in controlled settings, we now test whether the same pattern appears in standard NLP model families. These experiments are not architecture-controlled—BERT-style encoders and GPT-2-style autoregressive models differ in attention structure, parameterization (∼110M vs. ∼137M), and pretraining objective. We therefore interpret them as evidence about commonly used NLP pipelines rather than as standalone causal proof. The key question is whether the controlled-setting findings have ecological validity in text.

5.1. Stylized SST-2: Synthetic Text Shortcuts

We inject style-based shortcuts into SST-2 (Socher et al., 2013) by rewriting examples into GENZ or SHAKESPEAREAN style (using Claude Sonnet 3.5) while preserving sentiment. Positive examples are paired with GENZ and negative with SHAKESPEAREAN at varying correlation strengths (50/50 to 90/10).

Style Transfer Example

Original: "An edgy thriller that delivers a surprising punch."
GENZ: "Ngl this movie's lowkey intense and hits different with a plot twist you don't see coming fr."
SHAKESPEAREAN: "A tale most daring, 'tis a thrilling yarn that doth deliver a blow most unexpected."

At test time, styles are randomly assigned, breaking the training correlation. We define four groups by the cross-product of label $y \in \{0, 1\}$ and style $s \in \{\text{GENZ}, \text{SHAKESPEAREAN}\}$, and report worst-group accuracy $\min_{(y,s)} \text{Acc}(y, s)$.

This experiment tests reliance on a surface-level style cue, not whether one architecture is intrinsically better at recognizing a particular writing style. If GPT-2 were simply worse at Shakespearean text regardless of sentiment, it would hurt *both* positive-Shakespearean and negative-Shakespearean groups equally, and worst-group accuracy would not specifically collapse on the minority group where style conflicts with the training label. The diagnostic quantity is whether accuracy degrades *specifically* when style conflicts with sentiment—the signature of shortcut reliance.

Results (Figures 5 and 6). Pure-Gen GPT-2 exhibits substantially lower worst-group accuracy than BERT at every spurious correlation strength. At 90/10, GPT-2's worst-group accuracy drops to near zero while BERT maintains ∼0.27—despite comparable overall accuracy (∼0.90). The gap widens monotonically with correlation strength (Figure 5), consistent with the controlled-setting finding that generative objectives amplify reliance on available spurious cues. This pattern holds across all model sizes (Small, Medium, Large, Pretrained) and training data scales.

The BERT/GPT-2 model-family comparison is consistent with the controlled toy findings, but because these models differ in architecture and pretraining, this experiment should be interpreted as supporting ecological evidence rather than as an independent causal test.

5.2. CivilComments: Real-World Demographic Shortcuts

CivilComments (Koh et al., 2021) is a binary toxicity classification dataset with eight identity annotations (male, female, homosexual_gay_or_lesbian, christian, jewish, muslim, black, white) that induce naturally occurring spurious correlations. The dataset contains 269K/45K/134K train/val/test examples.

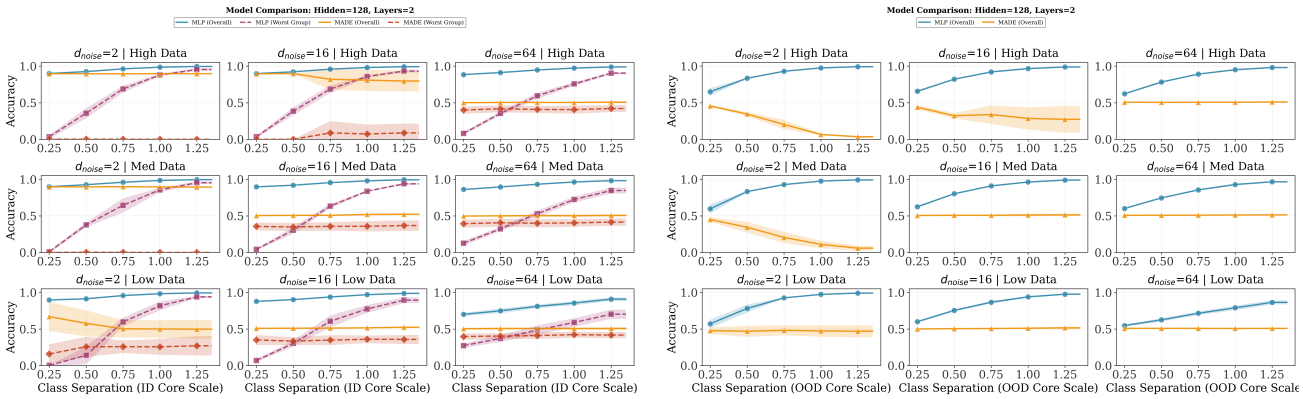
From Table 4, Pure-Gen GPT2 consistently exhibits worse worst-group accuracy than BERT across all model sizes, with the Small configuration achieving 0.000 worst-group accuracy (complete failure on the hardest demographic subgroup). The pseudo-generative variants substantially improve worst-group accuracy—Pseudo-Gen GPT2 achieves 0.529 and 0.540 at Medium and Large scales—suggesting that the classification interface (single forward pass vs. full sequence scoring) is a key determinant of shortcut robustness.

6. Discussion

Our tiered experimental design reveals a consistent pattern: in the controlled settings where architecture and capacity are matched, pure generative classification exhibits worse worst-group performance under spurious correlations. The same pattern then appears in NLP model-family comparisons, providing ecological validity.

The mechanism is intuitive: generative classifiers model $P(X \mid Y)$, which requires capturing *all* statistical regularities of the input conditioned on the label. When spurious features are strongly correlated with labels, they become part of the learned class-conditional density. At test time, examples where the spurious correlation is absent or reversed receive low likelihood, leading to misclassification. Discriminative classifiers only need to learn the decision boundary $P(Y \mid X)$ and can ignore features unnecessary for discrimination.

Pseudo-generative models emerge as a compelling middle ground: they are trained with generative objectives (capturing some benefits of language modeling) but classify via a single forward pass that restricts prediction to the label



(a) In-distribution *(b)* Out-of-distribution

Figure 1. Overall and worst-group accuracy across data sizes, noise dimensions, and core scale values for MADE and MLP. The backbone and training protocol are fixed; only the classification objective differs.

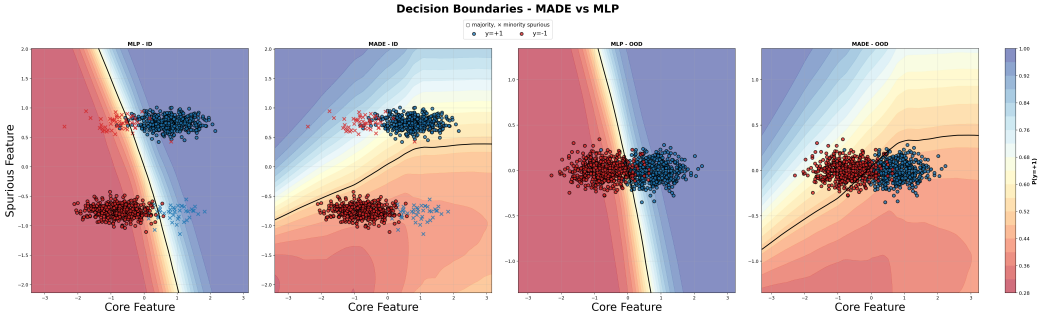


Figure 2. Decision boundaries for MLP and MADE under ID and OOD settings ($\mu_{\text{core}} = 0.75$). Architecture, capacity (∼6% difference), and training are matched; only the generative/discriminative objective changes. The MLP learns a stable linear boundary; MADE produces distorted contours that fail under distribution shift.

position, avoiding the need to score the full input under each candidate label. This architectural choice appears to reduce shortcut reliance while maintaining competitive overall accuracy.

Our results establish that generative classification is not automatically protected against shortcut learning; they do not imply that every generative model on every dataset will be worse. The effect is strongest when spurious features are highly available and predictive—precisely the conditions under which shortcut learning is most concerning in practice.

7. Conclusion

Across controlled Gaussian and tabular-transformer settings, pure generative classifiers were not inherently shortcut-averse: when spurious features were highly available, they achieved competitive average accuracy while degrading worst-group performance. Stylized SST-2 and CivilComments experiments show that the same pattern appears in realistic NLP model-family comparisons. Pseudo-generative variants often improve worst-group accuracy, suggesting that the classification interface—and the degree to which the model must explain all of X—are important determinants of shortcut reliance. Our results show that shortcut avoidance is not guaranteed by generative modeling alone and must be evaluated directly with group-sensitive metrics.

Limitations

Our experiments use a tiered design that separates controlled comparisons from model-family comparisons, and these have different evidential status. The Gaussian and tabular-transformer experiments isolate classifier formulation under matched data, capacity (within 6%), and training protocols—these support mechanistic claims. The BERT/GPT-2 comparisons are ecologically relevant but not architecture-controlled: the models differ in attention structure, pretraining objective, and parameter count (∼110M vs. ∼137M). These results demonstrate that the pattern appears in commonly used NLP model families, but they are not standalone causal proof.

The controlled toy experiments use $\rho = 0.9$ as a stress test. We do not claim the magnitude of the effect is unchanged

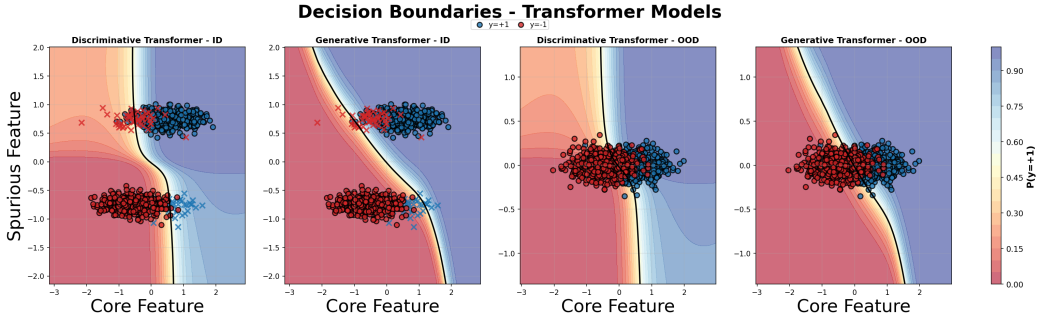
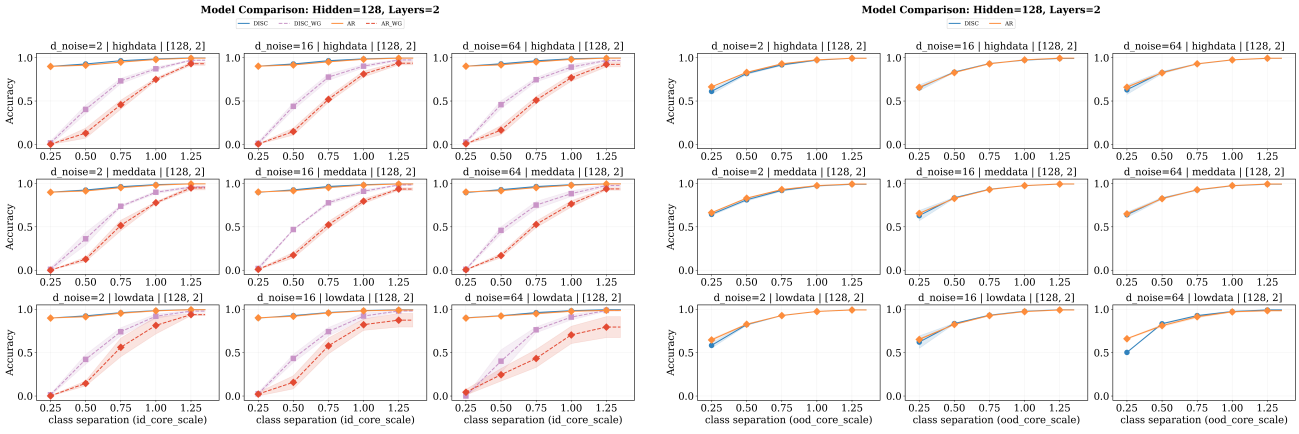


Figure 3. Decision boundaries for discriminative and autoregressive transformers ($\mu_{\text{core}} = 0.75$). The backbone is shared; only the attention mask and objective differ. The same pattern as MLP/MADE emerges.



(a) In-distribution *(b)* Out-of-distribution

Figure 4. Accuracy for discriminative and autoregressive transformer models. Data, backbone, and training are fixed; only the generative/discriminative formulation changes.

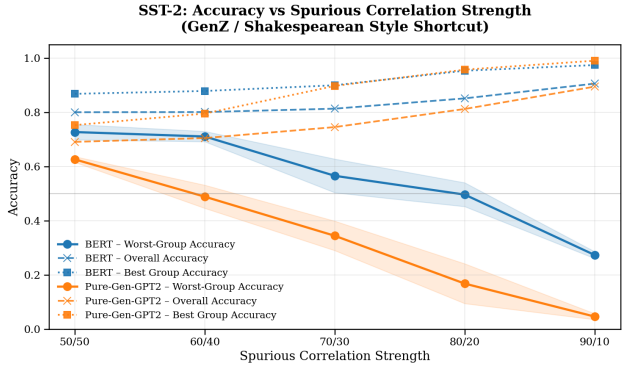


Figure 5. Worst-group accuracy vs. spurious correlation strength. The gap between BERT and Pure-Gen GPT-2 widens monotonically, consistent with the controlled-setting mechanism.

at lower spurious-correlation strengths; weaker correlations would likely attenuate the observed gap. However, the SST-2 ablation (Figure 5) shows the effect is monotonic across correlation strengths from 50/50 to 90/10, confirming that the mechanism operates across a range of shortcut intensities.

The stylized SST-2 setting uses GENZ and SHAKESPEAREAN style transformations as controllable surface shortcuts. A model-specific weakness on a particular style could affect absolute accuracy; however, our diagnostic is worst-group accuracy over the style×label cross-product, which isolates shortcut reliance from style competence. If a model were simply bad at one style regardless of sentiment, both groups sharing that style would suffer equally, and the minority-group-specific collapse we observe would not appear.

Finally, our real-world evaluation focuses on CivilComments and identity-based subgroup structure. The conclusions may not generalize to all forms of distribution shift, to ordinal classification (Kasa et al., 2024), to parameter-efficient fine-tuning (Hu et al., 2022), or to few-shot in-context learning. A unifying theory explaining when generative modeling reduces or amplifies shortcut reliance across modalities remains an important direction for future work.

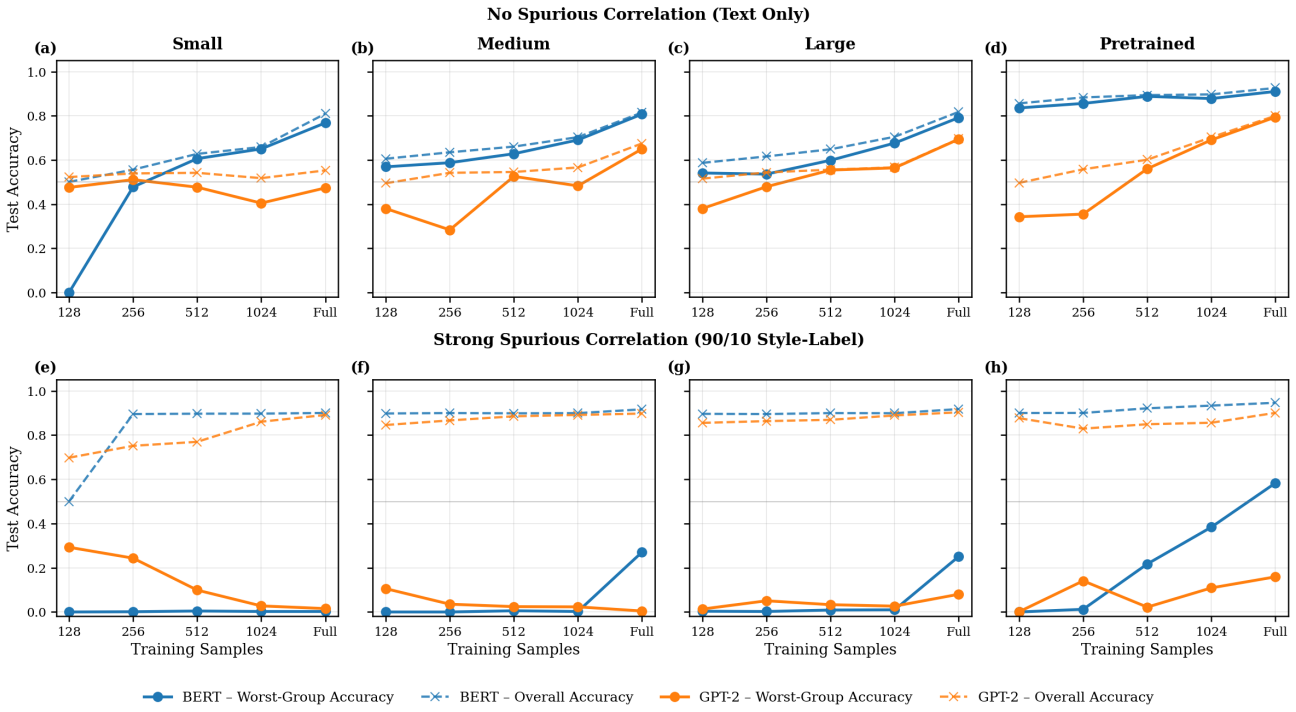


Figure 6. Accuracy under different model settings in SST-2. This is a model-family comparison (BERT vs. GPT-2); the training data, evaluation protocol, and shortcut injection are held fixed.

Table 4. CivilComments performance. Each cell: **Weighted F1 / Accuracy / Worst-Group Accuracy** (↑). In this model-family comparison, Pure-Gen GPT2 consistently exhibits behavior consistent with stronger reliance on the spurious demographic cue, particularly visible in worst-group accuracy.

Model Type	Small (3.3M)	Medium (30.3M)	Large (120.4M)	Pretrained
`BERT` (discriminative)	0.910 / 0.916 / 0.381	0.913 / 0.917 / 0.433	0.913 / 0.919 / 0.422	0.925 / 0.930 / 0.448
`Pure-Gen GPT2` (generative)	0.833 / 0.886 / 0.000	0.883 / 0.897 / 0.271	0.893 / 0.900 / 0.367	0.876 / 0.883 / 0.328
`Pseudo-Gen GPT2`	0.857 / 0.893 / 0.038	0.916 / 0.916 / 0.529	0.919 / 0.921 / 0.540	0.920 / 0.925 / 0.406
`Pseudo-Gen MLM`	0.833 / 0.886 / 0.000	0.921 / 0.923 / 0.483	0.919 / 0.921 / 0.442	0.923 / 0.925 / 0.500

Impact Statement

This paper presents work whose goal is to advance understanding of robustness in text classification. Our findings are directly relevant to building fair NLP systems: we demonstrate that generative classifiers can exhibit degraded worst-group performance that disproportionately affects minority subgroups, even when average accuracy appears competitive. We believe these insights can inform the design of more equitable AI systems.

References

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., and Love, B. C. A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, 166:164–171, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Du, M., He, F., Zou, N., Tao, D., and Hu, X. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2023.

Du, Y., Li, B., Torralba, A., Tenenbaum, J. B., and Isola, P. Shortcut learning of large language models in natural lan-

guage understanding. *arXiv preprint arXiv:2208.11857*, 2022.

Efron, B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898, 1975. doi: 10.1080/01621459.1975.10480319.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 881–889, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/germain15.html.

Hermann, K., Mobahi, H., Fel, T., and Mozer, M. C. On the foundations of shortcut learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Tj3xLVuE9f.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jaini, P., Clark, K., and Geirhos, R. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Kasa, S. R., Goel, A., Gupta, K., Roychowdhury, S., Priyatam, P., Bhanushali, A., and Srinivasa Murthy, P. Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5390–5404, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.320. URL https://aclanthology.org/2024.findings-acl.320/.

Kasa, S. R., Roychowdhury, S., Gupta, K., Kumar, A., Biruduraju, Y., Kasa, S. K., Pattisapu, N. P., Bhattacharya, A., Agarwal, S., and Huddar, V. Generative or discriminative? revisiting text classification in the era of transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8634–8648, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.486. URL https://aclanthology.org/2025.emnlp-main.486/.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Li, A. C., Kumar, A., and Pathak, D. Generative classifiers avoid shortcut solutions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Makroo, O., Kasa, S. R., Roychowdhury, S., Gupta, K., Pattisapu, N., Kasa, S., and Negi, S. The hidden cost of modeling P(X): Vulnerability to membership inference attacks in generative text classifiers. *arXiv preprint arXiv:2510.16122*, 2025. URL https://arxiv.org/abs/2510.16122.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Socher, R., Peres, A., Potts, C. D., Manning, C. D., and Wu, A. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642, 2013.

Stanley, E. A., Forkert, N. D., and Wilms, M. Does a diffusion-based generative classifier avoid shortcut learning in medical image analysis? an initial investigation using synthetic neuroimaging data. In *Medical Imaging 2025: Imaging Informatics*, volume 13411, pp. 94–99. SPIE, 2025.

Tu, L., Lalor, J. P., and Yu, H. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, A. and Cho, K. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36, 2019.

Xue, J.-H. and Titterington, D. M. Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural Processing Letters*, 28(3):169–187, 2008. doi: 10.1007/s11063-008-9088-7.

Yogatama, D., Dyer, C., Ling, W., and Blunsom, P. Generative and discriminative text classification with recurrent neural networks. In *arXiv preprint arXiv:1703.01898*, 2017.

Zheng, C., Wu, G., Bao, F., Cao, Y., Li, C., and Zhu, J. Revisiting discriminative vs. generative classifiers: Theory and implications. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42420–42477. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zheng23f.html.

A. Parameter Count Derivations

MLP. Architecture: $[D_{\text{in}}, H, H, 2]$. Parameters: $(D_{\text{in}} \cdot H + H) + (H^2 + H) + (2H + 2)$. For $D_{\text{in}} = 4$, $H = 128$: $4 \times 128 + 128 + 128^2 + 128 + 2 \times 128 + 2 = 17{,}410$.

MADE. Architecture: $[D_{\text{in}} + 2, H, H, 2D_{\text{in}}]$ with autoregressive masks. Parameters: $((D_{\text{in}} + 2) \cdot H + H) + (H^2 + H) + (2D_{\text{in}} \cdot H + 2D_{\text{in}})$. For $D_{\text{in}} = 4$, $H = 128$: $6 \times 128 + 128 + 128^2 + 128 + 8 \times 128 + 8 = 18{,}440$. Note that MADE's output layer scales with $2D_{\text{in}}$ (producing Gaussian μ, σ for each input dimension), so at higher noise dimensions the gap grows (e.g., $D_{\text{in}} = 66$: MLP = 25,346 vs. MADE = 42,372). However, the MLP's consistent worst-group advantage across *all* configurations—including those where MADE has substantially more parameters—confirms that the result is driven by the objective, not by capacity disadvantage.