When Personalization Tricks Detectors: The Feature-Inversion Trap in Machine-Generated Text Detection

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly imitate personal writing styles, personalization has become a key challenge for machine-generated text (MGT) detection. Yet personalized MGT detection remains largely underexplored. In this work, we introduce `StyloBench`, the first benchmark for evaluating detector robustness under personalization, built from literary and blog texts paired with their LLM-generated imitations. Experiments across diverse detectors show pronounced performance instability under personalization, with frequent inversions relative to general-domain behavior. To better understand this limitation, we conduct an in-depth analysis and attribute it to a *feature-inversion trap*, i.e., features that are effective for separating human-written text (HWT) from MGT in general flip their effect in personalized contexts, ultimately misleading detectors. Motivated by this, we propose `StyloCheck`, a diagnostic framework for predicting detector robustness under personalization. `StyloCheck` identifies the inverted features and quantifies detector dependence using perturbed texts pronounced in the features. In our experiments, `StyloCheck` predicts both the direction and magnitude of cross-domain performance shifts with an 85% correlation to actual outcomes. We hope this work will raise awareness of the structural risks introduced by personalization and motivate more robust approaches to personalized MGT detection. Github.[1]

1 Introduction

Large Language Models (LLMs) have achieved strong text generation performance (Huang et al., 2025), with increasing capability of mimicking personalized language styles in tasks such as news writing, style imitation, and story generation (Tu et al., 2024; Wang et al., 2025). However, these capabilities also raise security and ethical concerns,

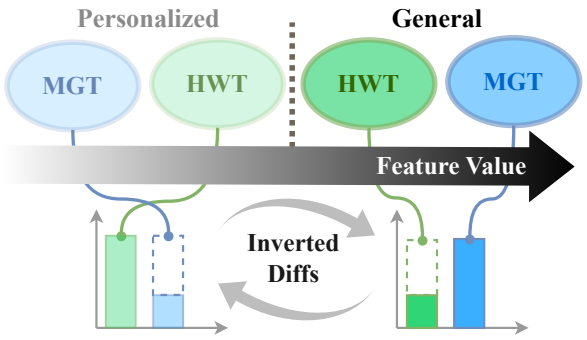

Figure 1: Illustration of the feature-inversion trap. The feature values of HWT/MGT exhibit inverted differences across domains.

as LLMs may generate fake news and misinformation (Tian et al., 2025). Moreover, style imitation can be misused, for instance, by impersonating public figures or creating fake work (Herbold et al., 2024a). These risks make machine-generated text (MGT) detection increasingly important. Although existing studies have made progress in general-domain detection, it remains unclear how well they perform in personalized domains. Hence, we introduce `StyloBench`, the first benchmark for MGT detection in personalized settings. `StyloBench` covers two sub-scenarios: Literary works and Blog posts, each paired with LLM-generated imitations. Interestingly, experiments on `StyloBench` show that personalization can noticeably degrade detector performance. Moreover, on the Literary subset, many detectors exhibit *prediction inversion*, where predictions tend to shift opposite to the expected direction, suggesting weakened or even flipped discriminative cues under highly personalized text.

In order to explore why detectors fluctuate under personalization, we train a domain classifier on human-written texts (HWT) from general and personalized domains, and test it on both HWT/MGT in the two domains to see, across domains, whether domain features fluctuate similarly. The classifier exhibits a clear inversion: In the general domain,

[1] Code is also available in the supplementary metarials.

MGT is predicted to be more "general" than HWT; however, in the personalized domain, MGT is instead classified as more "personalized" than HWT.

This has given rise to our *feature-inversion trap* hypothesis, namely that features that are effective for separating HWT from MGT in general flip their effect in personalized contexts, ultimately misleading detectors. In order to identify where this inversion is most pronounced, we formalize the search for the strongest inversion as a Rayleigh quotient problem (Dong et al., 2024) and exploit its extremal property to obtain the *inverted feature direction*. Feature directions derived from different datasets exhibit high cosine similarity, closely aligned with domain classifier weights and far from those of HWT/MGT classifiers. Projecting samples onto these feature directions yields scalar *feature values*, as shown in Figure 1, which reveal a clear inversion effect: within each dataset, the HWT–MGT feature value difference is negatively correlated with detector performance. This suggests that the "feature-inversion trap" is a stable, cross-domain phenomenon akin to stylistic differences, and that detector failures are partly driven by reliance on these inverted features.

Based on this finding, we propose `StyloCheck`, an effective approach to predict the detector's performance changes in personalized scenarios. Given a detector, `StyloCheck` evaluates it on probe datasets constructed with token-level perturbations that remove semantics, style, and basic HWT/MGT features while preserving inverted-feature differences. The resulting performance reflects the detector's reliance on inverted features: higher performance on probe datasets indicates stronger reliance on these features. We test seven detectors on 100 probe datasets and find that the Pearson correlation between `StyloCheck`'s outputs and the actual cross-domain performance gaps exceeds 0.7 in 90% of the cases, and consistently stays above 0.85. This shows that `StyloCheck` reliably predicts both the direction and the magnitude of transfer performance changes, with higher reliability as the number of probe datasets increases.

Our contributions are as follows: (1) We build `StyloBench`, the first benchmark for MGT detection in personalized scenarios, and uncover drastic performance declines and even reversals in existing detectors. (2) We identify the *Feature-Inversion Trap*, a systematic shift between general and stylized domains, and show that this phenomenon can fundamentally undermine detectors by inverting the features they rely on. (3) We propose `StyloCheck`, which predicts the trend and the magnitude of a detector's performance changes in personalized scenarios. It exhibits high reliability and strong consistency with actual performance, with Pearson correlation exceeding 0.85.

2 Related Work

2.1 MGT Detection

Several benchmarks have been developed for evaluating the performance of HWT vs. MGT detection across different domains, generators, and languages, e.g., MGTBench (He et al., 2024), `M4` (Wang et al., 2024b), `M4GT-Bench` (Wang et al., 2024a) and RAID (Dugan et al., 2024). Moreover, several domain-specific benchmarks, including WetBench (Quaremba et al., 2025) and MultiSocial (Macko et al., 2025), have focused on specialized contexts such as Wikipedia and social media. However, none of them has looked into evaluating MGT detection in highly personalized or stylistically consistent text.

MGT detection methods fall into two main categories: training-based and training-free (Xu et al., 2025a). Training-based methods treat detection as supervised classification, usually by fine-tuning pretrained encoders such as RoBERTa (Liu et al., 2019), or by using improved frameworks and architectures (Guo et al., 2024; Tian et al., 2024; Jiao et al., 2025). Training-free methods rely on explicit textual or probabilistic cues, including geometric and probabilistic signals (Bao et al., 2024; Xu et al., 2025a), token distributions (Su et al., 2023b), topological features (Tulchinskii et al., 2023; Wei et al., 2025), and human-assistive indicators (Gehrmann et al., 2019b; Russell et al., 2025). However, existing studies have not addressed or evaluated personalized or highly stylistically adaptive scenarios.

2.2 Personalized LLM Generation

Personalization of LLMs has recently become increasingly important (Zhang et al., 2025c). Personalization methods have developed into two main approaches: (*i*) *Prompt-based personalization* drives LLMs toward users' traits via personalized prompts (Tseng et al., 2024). Some work designed retrieval (Mysore et al., 2024) and agent frameworks (Zhang et al., 2025b) to achieve deeper imitation. (*ii*) *Training-based personalization* adapts user traits via instruction tuning (Woźniak et al., 2024; Liu et al., 2025), or through self-

Subset	Stylo-Literary	Stylo-Blog
Domain	Article	Blog
Generator Size	$\leq$ 14B	$\geq$ 70B
Method	CPT	Prompting
Generators	3	4
Subdomains	7	1
Examples	21,000	4,000
Sample Length	$\leq$ 512 tokens	$\leq$ 512 tokens

Table 1: Statistics about StyloBench.

supervised learning for dynamic adaptation (Mendoza et al., 2024). This often yields stronger and more persistent stylistic alignment. However, such personalization ability of LLMs raises concerns, including the possibility for political impersonation (Herbold et al., 2024b) and copyright infringement (Zhang et al., 2025a; Karamolegkou et al., 2023). This underscores the need for personalized MGT detection. However, to the best of our knowledge, no prior work has systematically studied the problem of personalized MGT.

3 StyloBench

To investigate the performance of existing MGT detection methods in personalized scenarios, we create StyloBench, the first benchmark for MGT detection under personalized conditions. This dataset has two subsets representing two scenarios: (*i*) Stylo-Literary, simulating personalization in literary works, and (*ii*) Stylo-Blog, simulating personalization in blogs. Table 1 gives some statistics about these datasets. We provide a detailed discussion about motivation for dataset scenario selection, model selection, and dataset scale and diversity in Appendix A.1.

3.1 Dataset Construction

3.1.1 Stylo-Literary

In the article scenario, HWTs consist of excerpts from literary works, while MGTs are generated by LLMs trained to learn and imitate the authors' styles. Concretely, for HWT, we use data from the Gutenberg Book Corpus (Gerlach and Font-Clos, 2020), an open-source collection of books grouped by authors. We selected seven authors: Jane Austen (J.A), Charles Dickens (C.D), Fyodor Dostoyevsky (F.D), Plato (P.L), Bernard Shaw (B.S), Jonathan Swift (J.S), and Mark Twain (M.T). These authors are well known for their distinctive styles and each has more than five long-form works. As some artifacts in the texts are not original content, but rather formatting or source information, we clean the texts to keep only the original content (see Appendix A.3 for more detail). Then, we split each author's texts into 512-token segments. For each author, we randomly select 1,000 segments as HWTs in the test set, and up to 3,000 additional segments as the training set. For MGT, we apply continuous pre-training (CPT, Shi et al., 2024) of LLMs on the training set to achieve deeper personalized imitation. We train three LLMs in their base versions: Qwen3-4B (Team, 2025), Llama-3.1-8B (Dubey et al., 2024), and Phi-4 (14B, Abdin et al., 2024). We update only one LoRA (Hu et al., 2022) weight to reduce the training cost and to speed up learning. After training, we take the first 30 tokens of each HWT test sample as the prompt and let the LLM continue the text. The selected hyperparameters and other generation details are in Appendix A.5.

3.1.2 Stylo-Blog

In the blog scenario, the HWTs come from Blog-1K (Wang, 2022), a high-quality subset of the Blog Authorship Corpus (Schler et al., 2006). Blog-1K contains multiple posts grouped by human authors. We further introduce the data source in Appendix A.4. We randomly select 1,000 posts, each truncated to a maximum length of 512 tokens, as HWT examples in the test set, and the corresponding MGTs are generated by LLMs. For each blog post, we apply a few-shot prompting template using 1–3 other posts by the same author as examples to guide the model in imitating the author's style. The generator continues from the first 30 tokens of the given post, producing text with approximately the same length as the original. We use four popular large-scale LLMs as generators: GPT-4o (OpenAI et al., 2024), Claude-4-Sonnet (Claude-4) (Anthropic, 2025b), Claude-3.7-Sonnet-Latest (Claude-3.7) (Anthropic, 2025a), and Qwen2.5-72B (Team, 2024). The full generation details are in Appendix A.5.

3.2 Evaluation Setup

Evaluation Datasets Apart from StyloBench for personalized scenarios, we also evaluate on an English subset of M4 (Wang et al., 2024b), to show the MGT detectors' performance in a general setup. The English subset of M4 spans diverse sources, and contains MGTs from four generators: ChatGPT, Cohere, text-davinci-003 (Davinci), and BLOOMz, with ∼3,000 MGTs per source–generator pair. We further explain the data source and give more statistics in Appendix A.6.

Detector \ Generator	M4(General)				Stylo-Blog				Stylo-Literary		
	Cohere	ChatGPT	Davinci	BLOOMZ	Qwen2.5-72B	Claude-4	Claude-3.7	GPT-4o	Llama3.1-8B	Phi-4	Qwen3-4B
Entropy	31.83	26.35	40.10	41.33	13.02	36.57	34.08	63.43	55.23	51.92	76.18
Lastde	97.69	97.48	83.70	88.03	92.69	68.58	50.96	6.67	69.88	65.67	62.57
Lastde++	98.22	98.67	84.41	80.91	99.07	83.37	88.27	58.41	60.38	47.57	39.78
Log-Likelihood	93.12	93.76	72.79	59.75	95.27	72.30	77.29	41.65	36.59	30.94	9.23
LogRank	94.21	94.94	73.18	69.42	95.61	71.31	75.37	34.89	38.53	32.60	10.44
Detect-LRR	94.88	96.13	74.03	84.34	94.85	67.16	66.99	20.26	45.73	40.62	19.43
Fast-DetectGPT	98.78	98.99	85.28	55.01	99.47	84.60	89.43	57.57	33.22	18.47	8.71
Avg.	86.96	86.62	73.36	68.40	84.28	69.13	68.91	40.41	48.51	41.11	32.33

Table 2: Performance of MGT detectors, grouped by generator: shown is AUROC given one generator. Blue: higher AUROC; green: lower AUROC.

Baselines & Evaluation We evaluate seven representative training-free detectors, which achieve strong performances in general domain MGT detection: Log-Likelihood (Solaiman et al., 2019), LogRank (Solaiman et al., 2019), DetectLRR (Su et al., 2023a), Entropy (Gehrmann et al., 2019a; Ippolito et al., 2020), Fast-DetectGPT (Bao et al., 2024), Lastde and Lastde++ (Xu et al., 2025a). Details are provided in Appendix B.1. We use AUROC as the evaluation metric following prior work (Xu et al., 2025a; Bao et al., 2024).

We mainly focus on training-free MGT detectors because they rely on a small set of explicit text features, making performance changes easier to interpret as feature shifts across domains. Training-based detectors learn more complex representations influenced by data and model factors, so their behavior is harder to attribute to specific textual properties. However, we also include experiments for training-based methods in Appendix B.3. We use AUROC to measure the overall performance of an MGT detector, as detailed in Appendix C.1.

3.3 Main Results

Table 2 presents the performance of various detectors across datasets. We report the average AUROC for each generator across all subdomains, with full experimental results detailed in Appendix B.2. The last row shows the average AUROC for each generator across all baselines. The results reveal four primary findings: (1) *Significant performance degradation* occurs in personalized settings. Average AUROC on `M4` (above 85%) falls sharply on stylized datasets, dropping to as low as 32.33% on `Stylo-Literary`—worse than random guessing. (2) *High variance under domain shift*: Detectors exhibit divergent trends; for instance, while Entropy improves from 31.83% in `M4` to 76.18% in `Stylo-Literary`, Lastde drops from 97.69% to 62.57%. (3) *Systematic and abnormal inversions*: Many detectors experience dramatic flips, with AUROC for methods like Fast-DetectGPT falling as low as 8.71%, indicating a near-complete reversal of discriminative capability. (4) *Increased instability in complex styles*: Fluctuations are more pronounced on `Stylo-Literary` than on `Stylo-Blog`, with abnormal reversals occurring more frequently. These observations suggest that existing MGT detectors may be highly unstable in personalized scenarios. When generators effectively imitate specific styles, detector performance tends to shift unpredictably or even reverse entirely. Detailed per-domain performance analyses are provided in Appendix D.1.

4 The Feature-Inversion Trap

Building on the observed instability in personalized scenarios, we next analyze its mechanism. In §4.1, we introduce the feature-inversion trap hypothesis; in §4.2, we extract the most salient inverted feature vector to verify this hypothesis; and in §4.3, we demonstrate its generality across datasets.

4.1 Feature-Inversion Trap Hypothesis

Probing Method To analyze the differences between HWT and MGT, we require a representation space that effectively captures semantic and stylistic features. Prior work has shown that the hidden space of pretrained language models often encodes diverse linguistic and stylistic properties along approximately linear directions (Mikolov et al., 2013). Based on this insight, we adopt GPT-2 (Brown et al., 2020) as a proxy model, using activations from its different modules as feature representations. We examine all modules of GPT-2, including attention layers, MLP layers, and residual streams. For each module, we extract the activation of the last token as the text representation and train a logistic regression classifier to distinguish HWT from

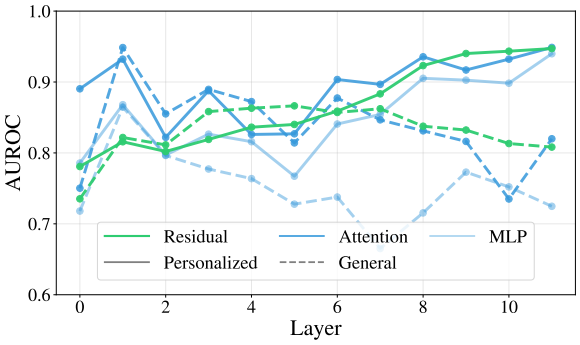

Figure 2: AUROC scores of MGT classifiers across modules in two domains. The colors denote modules, and the line styles denote domains. We can see that the deep residual layers are best at distinguishing HWT/MGT in both domains.

MGT in general and personalized domains. Figure 2 reports the AUROC for each module. Deep residual streams consistently achieve high AUROC across the two domains, indicating that they retain strong discriminative features. Therefore, we focus on the residual stream at a near-final layer, i.e., layer 10, in the following analysis.

Probe Datasets We construct two small probe datasets. Following (Xu et al., 2025a), we adapt Xsum (Narayan et al., 2018) and their LLM-generated continuations as the general-domain data. We randomly sample 150 texts from the J.A subset of `Stylo-Literary` and take their Phi-4 MGTs as personalized-domain data. Each domain contains 150 HWT and 150 MGT samples.

Visibility of the Inverting Trend Since the detector's performance fluctuates under domain shift, we first investigate whether the representations themselves encode domain differences. Following (Gao et al., 2025), we train a logistic regression-style domain classifier, whose weight direction serves as domain-related features, to distinguish general HWT from personalized HWT, and then evaluate it on HWT/MGT samples in both domains. Figure 3(a) shows a clear separation of the projection onto the weight direction (feature values) between the two domains, as expected. We also observe an unexpected pattern: in the general domain, the feature value of MGT is slightly lower than that of HWT, whereas in the personalized domain, it is slightly higher. This leads to the *feature-inversion trap hypothesis*: distinguishable MGT features in general domain are inverted under personalization.

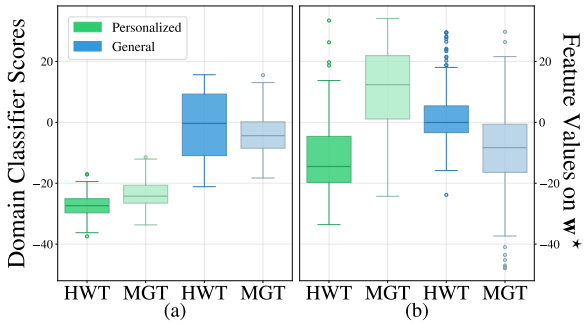

Figure 3: (a) Domain classifier score distribution: We can see moderate inversion effects. (b) Feature value distributions on the inverted feature direction $\mathbf{w}^\star$: we can see a major inversion effect.

4.2 Verification of the Feature-Inversion Trap

In order to verify the existence of the feature-inversion trap, we aim to identify the most representative inverted feature direction. If the projection of datasets in this direction is highly correlated with detector performance, it would suggest that detectors rely on it, thereby supporting our hypothesis.

4.2.1 Deriving the Inverted Feature Direction

We begin by extracting the inverted feature direction that is most responsible for this effect, in order to assess its correlation with detector performance.

Notation We denote the general-domain dataset by G and the personalized-domain dataset by S. Let $g_+, g_- \in G$ and $s_+, s_- \in S$ represent MGT and HWT activations in the two domains. For each quadruple (g_+, g_-, s_+, s_-), we compute domain-specific difference vectors:

$$v_G = g_+ - g_-, \quad v_S = s_+ - s_-. \tag{1}$$

Inversion-Value Matrix and Object Our goal is to find a direction $\mathbf{w}$ where the projection of v_G is opposite to that of v_S. For each quadruple (g_+, g_-, s_+, s_-), we define the projection product in direction $\mathbf{w}$ as

$$q_i(\mathbf{w}) = (\mathbf{w}^\top v_G)(\mathbf{w}^\top v_S) = \mathbf{w}^\top (v_G v_S^\top)\mathbf{w}. \tag{2}$$

Since for any matrix M, it holds that $\mathbf{w}^\top \mathrm{M}\mathbf{w} = \mathbf{w}^\top \frac{1}{2}(\mathrm{M} + \mathrm{M}^\top)\mathbf{w}$, Equation 2 can be rewritten as

$$q_i(\mathbf{w}) = \mathbf{w}^\top \tfrac{1}{2}\big(v_G v_S^\top + v_S v_G^\top\big)\mathbf{w}. \tag{3}$$

Further details on the calculation of Equation 2 and 3 are in Appendix C.3. For each quadruple, we define a cross-domain matrix

$$A_i = \tfrac{1}{2}\Big(v_G v_S^\top + v_S v_G^\top\Big), \tag{4}$$

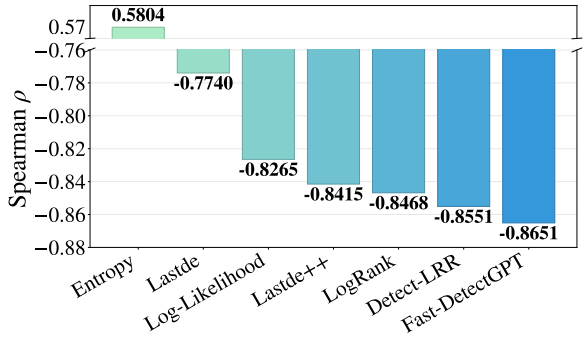

Figure 4: Spearman ρ between feature value differences of datasets and corresponding detector performance.

which is symmetry. Aggregating over all quadruples yields

$$A = \sum_i A_i. \tag{5}$$

The overall inversion objective can then be expressed as

$$\mathcal{R}(\mathbf{w}) = \sum_i q_i(\mathbf{w}) = \mathbf{w}^\top A \mathbf{w}, \quad \text{s.t. } ||\mathbf{w}|| = 1. \tag{6}$$

Since each A_i is symmetric, the aggregated matrix A in Equation 5 is also symmetric. So far, we have transformed the problem into the *Rayleigh quotient* of A with respect to $\mathbf{w}$ under the unit-norm constraint. Illustrations on the Rayleigh quotient problem are available in Appendix C.3.

Solution By the property of the Rayleigh quotient, minimizing the objective in Equation 6 reduces to: $\mathbf{w}^\star = \arg\min_{|\mathbf{w}|=1} \mathcal{R}(\mathbf{w})$, whose solution is the eigenvector of A corresponding to its smallest eigenvalue: $A = U\Sigma U^\top$, $\mathbf{w}^\star = U[:, -1]$, where $U[:, -1]$ denotes the last column of U, associated with the minimum eigenvalue. This $\mathbf{w}^\star$ represents the *inverted feature direction*, i.e., the axis along which the HWT–MGT projection (feature value) difference in the general domain is most strongly inverted in the personalized domain. Using the probe datasets in §4.1, we show the feature value distributions of four sample types on $\mathbf{w}^\star$ in Figure 3(b). In the personalized domain, the MGT feature values are clearly lower than for HWT, while in the general domain, the MGT ones are clearly higher than HWT's. This flip in relative positions provides direct evidence of the feature-inversion phenomenon.

4.2.2 Correlation with Detector Performance

Deriving the inverted feature direction $\mathbf{w}^\star$ reveals a dimension where MGT and HWT roles flip across domains, but this alone does not confirm its effect on detectors. To establish the connection, we evaluate the correlation between the strength of the inverted feature and detector performance.

Intuitively, along $\mathbf{w}^\star$, the relative positions of HWT and MGT feature values in the two domains are inverted. We quantify this property using *feature value difference*. For a dataset M with its MGT denoted as m_+ and HWT as m_-, the feature value difference is

$$\mathcal{D}(M, \mathbf{w}^\star) = \sum_{\{m_+, m_-\} \subset M} \left(m_+^\top \mathbf{w}^\star - m_-^\top \mathbf{w}^\star \right), \tag{7}$$

which reflects the overall discrepancy between MGTs and HWTs on the inverted feature. Larger values indicate a clearer separation, while smaller or flipped values suggest confusion between the two classes. Following the experimental design in §3.2, we partition `M4` and `StyloBench` by unique generator–subfield combinations, resulting in a total of $N = 45$ subsets. For each subset, we compute $\mathcal{D}(\cdot, \mathbf{w}^\star)$, forming a set $\{\mathcal{D}_i\}_{i=1}^N$. Meanwhile, for each MGT detector, we collect the AUROCs on the same subsets, denoted as $\{\text{AUROC}_i\}_{i=1}^N$. We measure their consistency by Spearman correlation: $\rho = \text{Spearman}\big(\{\mathcal{D}_i\}, \{\text{AUROC}_i\}\big)$. The resulting correlations ρ for each detector are shown in Figure 4. We can see that entropy exhibits a positive correlation (~ 0.6), whereas all other detectors have $\rho < -0.77$, indicating strong negative correlations. We further show the distribution of $\{\text{AUROC}_i\}$ versus $\{\mathcal{D}_i\}$ of each detector in Appendix D.2. Overall, these results demonstrate that the detector's performance is tightly linked to a feature inverted across domains.

To verify that this correlation is not spurious, we further conduct experiments that isolate the effects of inverted features. We evaluate detectors on randomized text lacking semantic content, where positive and negative samples are separated along the inverted direction, an orthogonal direction, or at random. Detectors only show strong discrimination under the first case, indicating their direct reliance on inverted features (Appendix D.3).

4.3 Generality of the Feature-Inversion Trap

Having verified that $\mathbf{w}^\star$ captures a key inverted feature correlated with performance, we now investigate whether this phenomenon is dataset-specific or reflects a broader, cross-domain pattern. To this end, we evaluate the consistency of inverted features across multiple datasets. We processed `Stylo-Literary` and `M4` as follows. For each

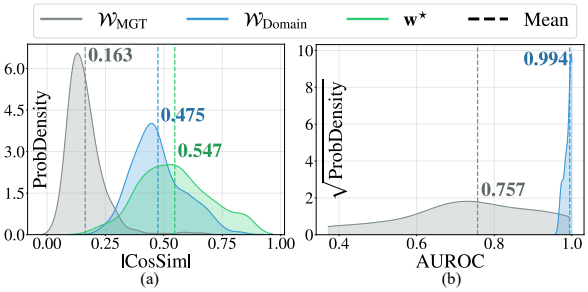

Figure 5: (a) |CosSim| between feature directions from different datasets. $\mathbf{w}^\star$ are close to $\mathcal{W}_{\text{Domain}}$ and beyond $\mathcal{W}_{\text{MGT}}$. (b) AUROC distributions of the generalization test. $\mathcal{W}_{\text{Domain}}$ has evidently better generalizability.

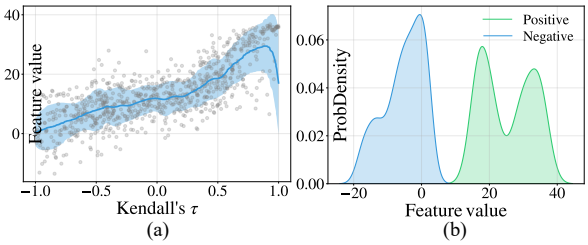

Figure 6: (a) Changes of feature values under different levels of token shuffling. Shuffling brings consistent change in feature value. (b) Distribution of feature values in the probe dataset. Two classes share zero overlap.

subdomain–generator pair in `M4`, we create five subsets, each with 150 random HWTs and 150 MGTs, as general-domain probe datasets. We apply the same procedure to `Stylo-Literary` to obtain personalized-domain probe datasets.

Each experiment samples one general subset and one personalized subset. We then extract three feature directions: (1) the inverted feature direction $\mathbf{w}^\star$, (2) the MGT feature direction $\mathcal{W}_{\text{MGT}}$, and (3) the domain feature direction $\mathcal{W}_{\text{Domain}}$. To obtain these two reference directions, we train logistic regression models. One model separates HWT and MGT and gives $\mathcal{W}_{\text{MGT}}$. The other separates general and personalized data and gives $\mathcal{W}_{\text{Domain}}$. We repeat this process 100 times and produce 100 sets of the three types of vectors. We then compute cosine similarity within each group. Figure 5(a) shows that $\mathcal{W}_{\text{MGT}}$ has low similarity with a mean of 0.163. $\mathcal{W}_{\text{Domain}}$ and $\mathbf{w}^\star$ show higher stability with means of 0.475 and 0.547. We also test how $\mathcal{W}_{\text{MGT}}$ and $\mathcal{W}_{\text{Domain}}$ generalize to new subsets. Figure 5(b) shows that $\mathcal{W}_{\text{Domain}}$ keeps a high AUROC in other subsets with a mean of 0.994, while $\mathcal{W}_{\text{MGT}}$ varies more widely from 0.4 to 0.8 with a mean of 0.757. The strong similarity between $\mathcal{W}_{\text{Domain}}$ and $\mathbf{w}^\star$ indicates that inverted features share the same high generalization ability.

Based on these observations, we conclude that the feature-inversion trap is a widespread phenomenon between personalized and general domains, and the inverted features share strong commonalities across various datasets.

5 StyloCheck

General MGT detectors suffer from the feature-inversion trap, leading to unreliable performance in personalized domains. In this section, we propose `StyloCheck`, an automatic transferability estimator that predicts such performance shifts by quantifying detectors' dependence on inverted features.

5.1 Design of StyloCheck

5.1.1 Probe Dataset Synthesis

To construct probe datasets that differ only in inverted features, we eliminate confounding factors from text semantics, domain, and class (HWT/MGT) by shuffling tokens. To control shuffle strength, we use Kendall's τ, a measure of sequence order ranging from 1 to -1, corresponding to a gradual inversion of token order (see Appendix C.2). For each sentence, we generate variants with Kendall's τ spanning this range. As shown in Figure 6(a), both Kendall's τ and the corresponding feature values vary continuously.

Therefore, we build probe datasets by shuffling tokens. We sample one general and one personalized HWT, generate 800 variants for each with different Kendall's τ, merge them, and select the 50 samples with the highest feature values as positives and the 50 lowest as negatives. As shown in Figure 6(b), the resulting feature value distributions show no overlap. We further evaluate the style and MGT linear classifiers introduced in §4.1, which achieve near-perfect accuracy during training but drop to 53% and 66% AUROC on the probe dataset, respectively, confirming the effective removal of domain and class features. The probe dataset thus reflects only differences in the inverted features.

5.1.2 Transferability Evaluation

We next describe how the detector performance on the probe dataset reveals its transferability. Our evaluated detectors treat MGTs as positive samples. In both the general domain and the probe dataset, positive samples tend to exhibit higher feature values. Therefore, if a detector relies on the inverted feature, it should perform well on the probe dataset. AUROC reflects the degree of reliance on inverted

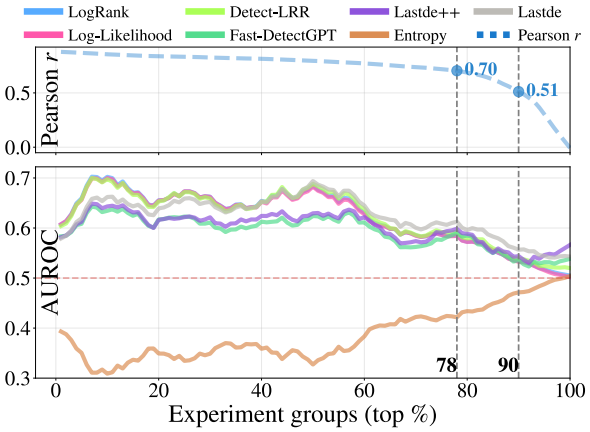

Figure 7: Top: Pearson r between transfer gaps and AUROCs. Bottom: corresponding AUROCs of detectors in probe datasets. Percentages of experiments groups where $r > 0.7$ and $r > 0.5$ are marked.

features: high values indicate strong dependence and likely degradation after transfer, low values (below 0.5) suggest inverted dependence and potential performance gains, and values near 0.5 imply weak dependence and stable transfer.

5.2 Performance of `StyloCheck`

To evaluate the performance of `StyloCheck`, we test it on all seven MGT detectors and examine whether it reflects their performance changes before and after transfer.

Evaluation Setup. We construct 100 probe datasets from `M4` and `Stylo-Literary`. In each experiment, we randomly choose five of them for testing, and we compute the mean AUROC. We then measure the Pearson r between this mean AUROC and the detector's overall performance gap between `M4` and `Stylo-Literary`. The results of over 100 such experiments are shown in Figure 7.

Results. Figure 7 shows the AUROC of all detectors across experiments. Two patterns appear. (1) Entropy stays below 0.5 in all runs, showing inverted reliance, while all other detectors remain above 0.5, showing positive reliance. This matches their transfer behavior, where Entropy improves, and others degrade. (2) In 90% of runs, Pearson r exceeds 0.5, and in 78% it exceeds 0.7, indicating stable reliance levels. These results show that `StyloCheck` identifies detectors' dependence on inverted features and also captures how strong that dependence is. We add an ablation study on the number of probe datasets in Appendix D.4.

6 Discussion

In this section, we address several conceptual questions raised by the observed mechanism, with empirical evidence deferred to Appendix E.

(i) Is the feature-inversion trap a typical out-of-distribution (OOD) effect? The feature-inversion trap is a special case of OOD, marked by two points: It aligns with the inverted feature direction, and it often causes a reversal of detector behavior rather than simple degradation. Appendix E.1 shows that common OOD do not produce these patterns. Appendix E.2 also outlines its links to related terms such as spurious correlations.

(ii) What do inverted features capture? Our evidence shows that they relate to text diversity. Many training-free detectors (Xu et al., 2025a; Gehrmann et al., 2019b) assume that HWT is more diverse than MGT, but personalization breaks this pattern: personalized MGT can be more varied and less coherent. This shift is consistent with the latent we observe and helps explain the negative performance flips of these detectors. See Appendix E.3.

(iii) How can we mitigate the feature-inversion trap? A practical option is to use tuned training-based detectors. They learn more cues and can reach strong in-domain accuracy after training on personalized text, though their cross-domain generalization remains limited, as shown in Appendix E.4. For training-free detectors, using features less sensitive to style drift, such as stable traits of human writing, may reduce reliance on diversity signals. Adaptive thresholding (Jung et al., 2025) that adjusts to the stylization can also improve robustness in personalized settings.

7 Conclusion and Future Work

We presented `StyloBench`, the first dataset for MGT detection in personalized scenarios. Our study showed that existing detectors face large performance shifts, and even inversion, after domain transfer. We traced this to the feature-inversion trap, where features that separate MGT and HWT change their roles across domains and lead detectors to flip predictions. Based on this, we proposed `StyloCheck`, a transferability framework that measures how much detectors rely on inverted features.

In future work, we plan to explore MGT detection methods that avoid such features to support stronger transferability.

Limitations

Our study primarily focuses on English, and further investigation is required to assess whether the observed findings generalize to multilingual, domain-specific, or code-switched settings. Linguistic variation across languages and domains may introduce distinct stylistic cues and distributional properties that affect both personalization and detection behavior in ways not captured by the current analysis.

While our results demonstrate that the feature-inversion mechanism plays a central role in explaining shifts in detector performance under personalization, other latent stylistic or semantic factors may also contribute to detection robustness. These factors, such as discourse structure, pragmatic intent, or higher-level narrative patterns, are not explicitly modeled in our framework and remain an open area for future exploration.

Finally, our experiments are conducted in controlled offline environments using static benchmarks, which may not fully reflect the dynamics of real-world deployment scenarios. In practice, personalized text generation and detection often occur in interactive and evolving contexts, including adaptive generation loops, human–AI coauthoring, and adversarial style imitation. Evaluating model behavior under these more realistic conditions could provide a more comprehensive understanding of robustness and generalization in practical applications.

Ethical Considerations

This work aims to advance understanding of machine-generated text (MGT) detection in personalized scenarios and is intended for research on transparency, robustness, and responsible AI use. All datasets used in our experiments are derived from publicly available sources, and no private, sensitive, or personally identifiable information is included. The generation process follows open and reproducible settings without targeting any real individuals. While our findings reveal potential weaknesses in existing detectors, they are presented to support the development of safer and more reliable detection systems rather than to facilitate misuse or impersonation. We encourage future research to apply these insights ethically, ensuring that detection technologies are used to mitigate misinformation and protect authorship integrity rather than to compromise it.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2025a. Claude 3.7 sonnet system card. Technical report.

Anthropic. 2025b. Claude opus 4 & claude sonnet 4 system card. Technical report.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *Proc. of ICLR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Xiangyu Dong, Xingyi Zhang, and Sibo Wang. 2024. Rayleigh quotient graph neural networks for graph-level anomaly detection. In *Proc. of ICLR*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proc. of ACL*.

Lang Gao, Kaiyang Wan, Wei Liu, Chenxi Wang, Zirui Song, Zixiang Xu, Yanbo Wang, Veselin Stoyanov, and Xiuying Chen. 2025. Evaluate bias without manual test sets: A concept representation perspective for llms. *Preprint*, arXiv:2505.15524.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019a. GLTR: Statistical detection and visualization of generated text. In *Proc. of ACL*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019b. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. *Proc. of NeurIPS*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*.

Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024a. Large language models can impersonate politicians and other public figures. *arXiv preprint arXiv:2407.12855*.

Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024b. Large language models can impersonate politicians and other public figures. *Preprint*, arXiv:2407.12855.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. of ICLR*.

Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, and 1 others. 2025. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Proc. of NeurIPS*, pages 125–136.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proc. of ACL*.

Kaijie Jiao, Quan Wang, Licheng Zhang, Zikang Guo, and Zhendong Mao. 2025. M-RangeDetector: Enhancing generalization in machine-generated text detection through multi-range attention masks. In *Proc. of ACL Findings*.

Minseok Jung, Cynthia Fuertes Panizo, Liam Dugan, Yi R., Fung, Pin-Yu Chen, and Paul Pu Liang. 2025. Group-adaptive threshold optimization for robust ai-generated text detection. *Preprint*, arXiv:2502.04528.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proc. of EMNLP*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proc. of ACL*.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. LLMs + persona-plug = personalized LLMs. In *Proc. of ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Dominik Macko, Jakub Kopal, Róbert Móro, and Ivan Srba. 2025. Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts. In *Proc. of ACL*.

Andrea Cristina McGlinchey and Peter J Barclay. 2024. Using machine learning to distinguish human-written from machine-generated creative fiction. *Preprint*, arXiv:2412.15253.

Rafael Mendoza, Isabella Cruz, Richard Liu, Aarav Deshmukh, David Williams, Jesscia Peng, and Rohan Iyer. 2024. Adaptive self-supervised learning strategies for dynamic on-device llm personalization. *Preprint*, arXiv:2409.16973.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proc. of AACL*.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proc. of EMNLP*, pages 12448–12460.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proc. of EMNLP*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex

Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Gerrit Quaremba, Elizabeth Black, Denny Vrandečić, and Elena Simperl. 2025. Wetbench: A benchmark for detecting task-specific machine-generated text on wikipedia. *Preprint*, arXiv:2507.03373.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of EMNLP*, pages 3980–3990.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In *Proc. of ACL*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *Proc. of AAAI*.

Benjamin Schweinhart. 2021. Persistent homology and the upper box dimension. *Discret. Comput. Geom.*, pages 331–364.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023a. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Proc. of EMNLP Findings*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023b. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Proc. of EMNLP*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. *EMNLP*.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, QINGHUA ZHANG, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale positive-unlabeled detection of AI-generated texts. In *Proc. of ICLR*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Proc. of EMNLP Findings*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Characterereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proc. of ACL*.

Eduard Tulchinskii, Kristian Kuznetsov, Kushnareva Laida, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of AI-generated texts. In *Proc. of NeurIPS*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Haining Wang. 2022. Blog-1k.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proc. of ACL*, pages 3964–3992.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proc. of EACL*.

Zixiao Wang, Duzhen Zhang, Ishita Agrawal, Shen Gao, Le Song, and Xiuying Chen. 2025. Beyond profile: From surface-level facts to deep persona simulation in llms. *ACL findings*.

Dongjun Wei, Minjia Mao, Xiao Fang, and Michael Chau. 2025. Short-PHD: Detecting short LLM-generated text with topological data analysis after off-topic content insertion. In *Second Conference on Language Modeling*.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *Proc. of ICDM*.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. 2025a. Training-free LLM-generated text detection by mining token probability sequences. In *Proc. of ICLR*.

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025b. Personalized generation in large model era: A survey. In *Proc. of ACL*, pages 24607–24649.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *CoRR*, abs/2402.12715.

Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. 2025a. LLMs and copyright risks: Benchmarks and mitigation approaches. In *Proc. of ACL*.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, Xiaoman Pan, Lian Xiong, Jingguo Liu, Philip S. Yu, and Xian Li. 2025b. Personaagent: When large language model agents meet personalization at test time. *Preprint*, arXiv:2506.06254.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025c. Personalization of large language models: A survey. *Transactions on Machine Learning Research*.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S. Jaakkola, Geoffrey J. Gordon, and Pradeep Ravikumar. 2022. Fundamental limits and tradeoffs in invariant representation learning. *J. Mach. Learn. Res.*, pages 340:1–340:49.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2023. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 4396–4415.

A Details on StyloBench and M4

A.1 Dataset Construction Discussion

A.1.1 Scenario Selection

StyloBench focuses on two personalization scenarios: literary style imitation and social media style blog generation. These scenarios are widely used in recent personalization studies (Xu et al., 2025b; McGlinchey and Barclay, 2024) and provide rich individual expression. Literary texts offer long narratives with stable author styles, while blog posts capture informal and self-expressive writing. Together, they form a practical testbed for personalized MGT, where detectors face realistic stylistic shifts when distinguishing MGT from HWT. The current version of StyloBench does not yet include multilingual or other specialized domains, and future extensions are planned to cover a broader range of personalized scenarios.

A.1.2 Model Selection

The 2 components of StyloBench use different model configurations, mainly due to differences in data availability and personalization depth. For literary data in Stylo-Literary, there are many texts per author, which makes CPT on each author feasible and allows deeper stylistic imitation. Smaller models are used in this setting because they are easier to train and adapt to individual authors. For blogs in Stylo-Blog, each user has fewer posts, so CPT is not suitable. In this case, large instruction-tuned models accessed through APIs are prompted to mimic user style, which captures shallower but flexible personalization. Stylo-Literary and Stylo-Blog, therefore, represent two complementary levels of personalization and together provide a diverse evaluation space for detectors.

A.1.3 Dataset Scale and Diversity

The overall scale of StyloBench is comparable to existing MGT detection benchmarks such as CH3-English (Guo et al., 2023), with about 25,000 versus 26,000 samples. StyloBench contains 21 subsets, which is similar to the number of subsets in M4 with 20 subsets, and each experimental setting uses more than 1,000 test samples. This design provides enough data for stable statistical analysis while keeping the benchmark manageable. In Stylo-Literary, seven authors with clearly different styles are included, and all of them lead to strong performance changes in detectors on HWT

and MGT. This indicates that the current author set is already informative, while also leaving room for future expansion to more writers and domains.

A.2 Examples of StyloBench

For an intuitive understanding, we show examples of Stylo-Blog and Stylo-Literary with different generators and sub-domains in Table 15 and 16.

A.3 Preprocessing for Stylo-Literary

Due to formatting and compilation issues, some artifacts in the Gutenberg Book Corpus are not part of the original texts and may distort detector performance. We clean all artifacts by: (1) remove indentation symbols not present in the original texts; (2) delete isolated line breaks and reduce multiple consecutive line breaks to a single one, since isolated breaks are used for line-width alignment and multiple breaks denote paragraph or chapter boundaries; (3) remove lines consisting of repeated "=" symbols, which usually mark chapter or section starts. Moreover, we remove unrelated segments not written by target authors: (1) delete lines containing links, which are often source annotations; (2) delete lines with long digit sequences, which usually indicate compiler contact information.

A.4 Data Source of Stylo-Blog

The data source of Stylo-Blog is Blog-1K (Wang, 2022), a subset of the Blog Authorship Corpus (Schler et al., 2006). The Blog Authorship Corpus was collected from posts of blogger.com in 2004. It spans a wide range of topics, is linguistically diverse, and has been widely used for studying stylistic and demographic features. On this basis, Blog-1K (Wang, 2022) was built by filtering out high-quality blogs and grouping them by authors, resulting in a collection of 1,000 authors and 16,132 blogs. Blog-1K ensures clearer author attribution, balanced data distribution, and consistent prose quality, which can be directly leveraged to construct Stylo-Blog for evaluating detectors in personalized domains.

A.5 Configurations of LLM Generators

In constructing Stylo-Literary, the key parameter settings for training the generator LLM are shown in Table 3. For both subsets of StyloBench, we apply the same generation settings: (1) the maximum generation length is 512 tokens, to ensure consistency of imitated styles, as both training data and few-shot samples are no longer than 512 tokens; (2) the temperature is set to 1, to avoid repetition in base LLMs (Li et al., 2023), and encourage LLMs to produce vivid and diverse personalized content. For all other parameters, we adopt their default settings. For Stylo-Blog, the prompt template is shown in Figure 8.

Parameter	**Value**
Maximum samples	3,000
Batch size	8
Learning rate	0.0001
Epochs	5
LoRA configuration	
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.1

Table 3: Key hyperparameters in CPT for generating Stylo-Literary MGTs.

Generator	**Wikipedia**	**Reddit**	**WikiHow**	**PeerRead**	**arXiv**	**Total**
Davinci	3,000	3,000	3,000	2,323	3,000	14,323
ChatGPT	2,995	3,000	3,000	2,344	3,000	14,339
Cohere	2,336	1,220	2,999	1,702	3,000	11,257
BLOOMz	3,000	3,000	3,000	2,340	3,000	14,340

Table 4: Statistics of the selected English subset of M4, where the values indicate the numbers of HWTs and MGTs, which are equal.

A.6 the English Subset of M4

The selected English subset contains 5 data sources to reflect diversity and daily language use: Reddit, Wikipedia, WikiHow, arXiv, and PeerRead. Each LLM generator uses 2–8 distinct prompts to produce varied MGTs. We select only the generators that have data in all five sources in the released version of M4. Table 4 reports the detailed statistics.

B Experiment Details

B.1 Baselines

We follow the same set of baseline detectors as in (Xu et al., 2025a). The implementation details of the seven baselines in this study are as follows.

Log-Likelihood (Solaiman et al., 2019). The average log probability of all tokens in a text is used as the metric. Texts with lower average likelihood are more likely to be MGT.

LogRank (Solaiman et al., 2019). The average log rank of tokens in the text, where ranks are determined by GPT-J's predicted probabilities, is used as the metric. Texts with higher ranks are more likely to be MGT.

```
def Prompt(historical_examples, target_request):
    return f"""Given a **BLOG REQUEST** from a USER to continue writing a blog, write a **BLOG POST**
mimicking the USER to satisfy the REQUEST.
Use the following instructions for your response:

1. You should maintain consistency in tone and style with the USER's historical blog posts.
2. You should imitate the language style of the USER's historical blog posts.
3. You should employ similar rhetorical methods as the USER's historical blog posts.
4. You must continue the BLOG POST for **at least 512 tokens**, expanding naturally on the REQUEST.

Here are some historical blog posts by the USER:
{historical_examples}

REQUEST (blog beginning, first ~30 words):
{target_request}

Write the BLOG POST to **continue the REQUEST**, mimicking the tone, style, and rhetorical methods of the
USER's historical blog posts."""
```

Figure 8: Prompt template for MGT in synthesizing `Stylo-Blog`. It takes 1–3 blogs from the same author as `historical_examples` and uses the first 30 tokens of the current blog as `target_request` for continuation.

Entropy (Gehrmann et al., 2019a). The average entropy of the predicted token distribution is computed. Texts with higher entropy values are more likely to be MGT.

DetectLRR (Su et al., 2023a). The ratio of log-likelihood to log-rank is taken as the score. Larger ratios indicate a higher chance of MGT.

Fast-DetectGPT (Bao et al., 2024). This method perturbs the input text to create contrast samples and compares the scoring differences between the original and perturbed texts. Larger discrepancies indicate that the original text is more likely to be MGT.

Lastde (Xu et al., 2025a). This method analyzes local and global diversity of token probability sequences, and combines them with likelihood information. Lower diversity relative to likelihood is more indicative of MGT.

Lastde++ (Xu et al., 2025a). It is an enhanced version of Lastde that normalizes scores using multiple contrast samples, making the results more stable. Higher normalized values suggest MGT.

Here we test all baselines under a black-box scenario, i.e., we cannot access generators when detecting MGTs. We use GPT-J-6B (Wang and Komatsuzaki, 2021) as a proxy model for any necessary information for detectors. Specifically, for Log-Likelihood, LogRank, Entropy, and DetectLRR, we compute statistical features by aggregating the probabilities and ranks predicted by GPT-J-6B at each token position; for DetectGPT, Fast-DetectGPT, Lastde, and Lastde++, we generate perturbed or sampled variants of the text and then compare GPT-J-6B's scoring results between the original and the contrast samples.

B.2 Full Results on Training-free Methods

Since full results on `Stylo-Blog` have been presented in Table 2, here we report the AUROC of each MGT detector across all subdomains and generators on `Stylo-Literary` in Table 10, and full results on `M4` in Table 11.

B.3 Experiments on Training-based Methods

Setup. We evaluate two representative training-based detectors: Roberta (Liu et al., 2019) and DeTeCtive (Guo et al., 2024). For `M4`, each subset is split by randomly selecting 1,000 HWT and MGT samples as the test set, with all remaining samples used for training. The finetuning configuration for Roberta is shown in Table 5. For DeTeCtive, we use the publicly released checkpoint that achieves the best performance on `M4`[2] and evaluate it on the same test sets. We use AUROC for evaluation.

Hyperparameter	Value
Batch size	32
Epochs	10
Learning rate	2×10^{-5}
Warmup steps	2000
Random seed	42

Table 5: Training configuration for RoBERTa.

[2] https://huggingface.co/heyongxin233/DeTeCtive/blob/main/M4_monolingual_best.pth

Results and Analysis. The averaged results on the three datasets are reported in Table 6. Both detectors perform well on M4, but on StyloBench their AUROC scores mostly fall in the 0.4–0.6 range, close to random prediction. Roberta shows near-random performance in about 64% of the personalized settings, while for DeTeCtive this proportion reaches 92%. Such frequent and structural performance drops are uncommon in standard domain generalization, suggesting that inverted feature behavior plays a significant role. At the same time, these training-based models do not collapse to extremely low AUROC values as some training-free detectors do, which indicates that they may rely on a broader set of learned features that prevent complete failure. Full results are provided in Tables 12–14.

Dataset	Roberta	DeTeCtive
Stylo-Literary	53.33	53.78
Stylo-Blog	64.77	49.73
M4	99.94	84.69
AUROC 0.4–0.6	64%	92%

Table 6: Average AUROC of training-based detectors.

C Metrics and Mathematical Tools

C.1 AUROC

AUROC measures detector performance across the full range of thresholds rather than relying on a fixed one, therefore reflecting the overall ability of the detector. AUROC ranges from 0.0 to 1.0, where 0.5 corresponds to random guessing, and 1.0 indicates perfect discrimination. It can be interpreted as the probability that a randomly selected machine-generated text is assigned a higher detection score than a randomly selected human-written text. Values below 0.5 indicate performance worse than random guessing and imply that the predictions are systematically inverted.

C.2 Correlation Coefficients

Spearman ρ Spearman's rank correlation coefficient measures the monotonic relationship between two variables by computing the Pearson correlation on their rank values. It is defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{8}$$

where d_i is the rank difference of the i-th sample. Values close to 1 or -1 indicate strong positive or negative monotonic correlation.

Pearson r Pearson's correlation coefficient captures the linear relationship between two variables. It is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{9}$$

Values near 1 indicate strong positive linear correlation, values near -1 strong negative correlation, and 0 denotes no linear correlation.

Kendall's τ Kendall's τ measures the ordinal association between two variables based on the number of concordant and discordant pairs:

$$\tau = \frac{(\text{\#concordant pairs}) - (\text{\#discordant pairs})}{\binom{n}{2}}. \tag{10}$$

In this work, τ is also employed as a quantitative measure of word order. Values close to 1 indicate strong agreement with the original order, values close to -1 indicate reversed order, and values around 0 correspond to randomized tokens.

C.3 Mathematical Design Explanation

Derivation of Equation 2 The equation $q_i(\mathbf{w}) = (\mathbf{w}^\top v_G)(\mathbf{w}^\top v_S)$ follows from basic matrix multiplication:

$$(\mathbf{w}^\top v_G)(\mathbf{w}^\top v_S) = (\mathbf{w}^\top v_G)(v_S^\top \mathbf{w}) = \mathbf{w}^\top (v_G v_S^\top)\mathbf{w}.$$

Derivation of Equation 3 For any real matrix $\mathrm{M} \in \mathbb{R}^{n\times n}$ and vector $\mathbf{w}$, the quadratic form equals its transpose, as it is a scalar:

$$(\mathbf{w}^\top M \mathbf{w})^\top = \mathbf{w}^\top M^\top (\mathbf{w}^\top)^\top = \mathbf{w}^\top M^\top \mathbf{w}.$$

Therefore

$$\mathbf{w}^\top \mathrm{M}\mathbf{w} = \tfrac{1}{2}\mathbf{w}^\top(\mathrm{M} + \mathrm{M}^\top)\mathbf{w}.$$

Applying this to $\mathrm{M} = v_G v_S^\top$ gives

$$q_i(\mathbf{w}) = \mathbf{w}^\top \tfrac{1}{2}(v_G v_S^\top + v_S v_G^\top)\mathbf{w}.$$

Rayleigh Quotient For a matrix $\mathrm{M} \in \mathbb{R}^{n\times n}$ and a nonzero vector $w \in \mathbb{R}^n$, the Rayleigh quotient is defined as

$$R(w) = \frac{w^\top \mathrm{M} w}{w^\top w}. \tag{11}$$

When $||w|| = 1$, this can be also written as: $R(w) = w^\top \mathrm{M} w$. When M is symmetric, $R(w)$ is bounded between the smallest and largest eigenvalues of M. In particular, the minimum of $R(w)$ equals the smallest eigenvalue and the maximum equals the largest eigenvalue, attained when w is the corresponding eigenvector.

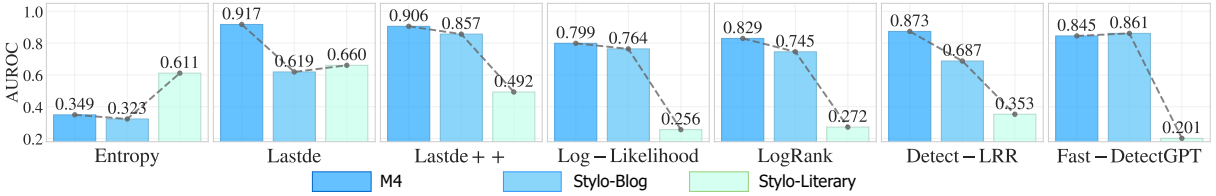

Figure 9: The average AUROC of different detectors on `M4`, `Stylo-Blog`, and `Stylo-Literary`. The lines indicate performance changes caused by domain transfer. Detectors show a clear change in the personalized domain, including surges, decreases, and inversions.

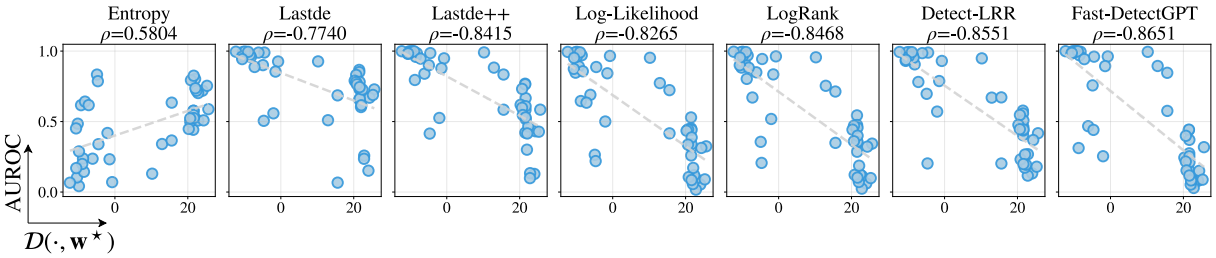

Figure 10: AUROC of each detector and the distribution of feature value difference $\mathcal{D}(\cdot, \mathbf{w}^\star)$ on the test set for each experiment. The plots show clear correlation, with varying strength and sign across detectors.

D Analysis on Feature-Inversion Trap

D.1 Performance Change of Detectors

To observe how each detector changes across the two domains, we compute its average AUROC on `M4`, `Stylo-Blog`, and `Stylo-Literary`, and plot the results in Figure 9. The figure shows results similar to §5.2: (1) Most detectors drop more in the personalized domain than in the general domain, while a few detectors, such as Entropy and Fast-DetectGPT, improve on some subsets of the personalized domain. (2) Most classifiers fluctuate less on `Stylo-Blog` than on `Stylo-Literary`, which indicates that deep personalized imitation after training misleads detectors more strongly than prompting. (3) On `Stylo-Literary`, five detectors fall below 0.5 in average AUROC, which shows that most detectors flip predictions in the personalized domain and confirms the generality of this phenomenon.

D.2 Correlation Visualization

Following §4.2.2, we plot the AUROC of each detector and the distribution of $\mathcal{D}(\cdot, \mathbf{w}^\star)$ on the test set across multiple experiments, as shown in Figure 10. The results reveal that: (1) The sample distribution is concentrated, and most points lie near the fitted line, which indicates a clear correlation between AUROC and $\mathcal{D}(\cdot, \mathbf{w}^\star)$. (2) Entropy shows a positive correlation, while the others show a negative correlation, which indicates that the Feature-Inversion Trap affects different detectors in different ways.

D.3 Verification of Detector Dependence on Inverted Features

In §4.2.2, we show a strong negative correlation between the projection gap of inverted features and detector performance. This correlation may still be spurious. We therefore design experiments to show that detectors do rely on inverted features. The core idea is simple: if a detector relies on inverted features, then it should still work on random text. The text may be meaningless, but if positive and negative samples show a clear gap along the inverted feature, the detector should still show strong results. On the contrary, other features or pure random noise should not produce this effect.

Setup. We tokenize each text. We then shuffle these tokens to create text with no meaning. Based on these texts, we build three types of test sets. (1) *Along inverted direction*. We sort samples by their projection on the inverted feature direction $\mathbf{w}^\star$. We take the top samples as positives and the bottom samples as negatives. This set tests whether the detector depends only on inverted features. (2) *Along orthogonal direction*. We sort samples by their projection on a direction $\mathbf{w}^\star_\perp$ that is orthogonal to the inverted feature. We take the top samples as positives and the bottom samples as negatives. This set tests whether the detector depends on other directions that share the same form but do not re-

flect inverted features. (3) *Random*. We shuffle all samples and split them into positives and negatives at random. This gives a baseline under complete randomness. For each type, we create 50 datasets. Each has 50 positive and 50 negative samples.

Results and analysis. We test seven detectors on these datasets. We plot the AUROC distribution across the 50 runs in Figure 11. The results show three clear patterns. (1) *Random*: All detectors stay near 0.5. This confirms that detectors behave randomly in random noise. (2) *Inverted direction*: Detectors show strong classification tendencies. Entropy has an average AUROC below 0.4. Other detectors have an AUROC above 0.6. The AUROC distribution is also more spread out. Extreme values such as AUROC above 0.8 or below 0.2 appear more often. These signs show that detectors rely on inverted features to make a prediction. (3) *Orthogonal direction*: No strong tendency appears. Some detectors behave almost the same as in the random set, such as Lastde, Lastde++, and Fast-DetectGPT. Other detectors shift slightly but remain within the 0.4 to 0.6 range, which is still close to random. These results show that other features cannot influence detector behavior as inverted features do. Together with the strong correlation shown in Figure 4, these results support a clear conclusion. In personalized settings, the feature-inversion trap is a key reason for the reversed behavior of many detectors.

D.4 Ablation Study for `StyloCheck`

We conduct an ablation study to assess how the number of probe datasets affects the reliability of `StyloCheck`. Figure 12(a) shows the distribution of r when using 1, 3, and 5 datasets. With fewer datasets, the mean r decreases and the probability of $r < -0.5$ increases, indicating a higher risk of incorrect prediction. Figure 12(b) plots the change of r as the number of datasets grows from 1 to 10. The mean r rises gradually, but slows down as it approaches an upper bound near 0.8. The standard deviation also decreases, but with diminishing returns. These results suggest that increasing the dataset size improves reliability, but the benefit diminishes over time. Using five probe datasets offers a good balance in practice.

E Empirical Evidence for §6

E.1 Difference Between the Feature-Inversion Trap and OOD Effects

In §6, we state two reasons why the feature-inversion trap is different from standard OOD effects: (1) it is tightly correlated with the inverted feature direction; (2) it causes reversal of detector behavior, not only performance decay. We now verify both points with experiments.

For point (1). We use performance differences across M4 subsets as a representative OOD case, since prior work (Wang et al., 2024b) uses them to study detector generalization and their domain gaps are well known. For all M4 subsets, we follow §4.2 and compute the projection gap on the inverted direction, denoted as $\mathcal{D}_{\text{M4}}$. We then compute the Spearman correlation between this projection gap and detector performance on M4:

$$r = \text{Spearman}(\mathcal{D}_{\text{M4}}, \text{AUROC}_{\text{M4}}). \tag{12}$$

The results for all detectors are shown in Figure 13. Cross-domain correlations (general → personalized) are consistently higher than correlations within M4 subsets. This shows that the inverted direction explains changes when moving to personalized domains, but not the differences inside M4.

For point (2). Reversal rarely appears across M4 subsets, but it occurs frequently on StyloBench. Table 11 reports all results on M4. Among 140 runs, only 25 (17.9%) have an AUROC below 0.5. In contrast, in Stylo-Literary, 93 out of 147 runs (63.3%) fall below 0.5. This large gap shows that personalized scenarios cannot be fully explained by common OOD factors such as feature weakening. Together with point (1), this supports that personalization causes a structural reversal of the same feature, not a typical domain shift.

E.2 Connections to Other Terms

The feature inversion phenomenon introduced appears to share surface similarities with several established concepts in machine learning. To avoid misunderstanding, we describe the relation between these concepts and our findings, and we clarify the differences.

Spurious Correlations. The feature inversion phenomenon is related to spurious correlations (Ye et al., 2024) because many detectors rely on cues

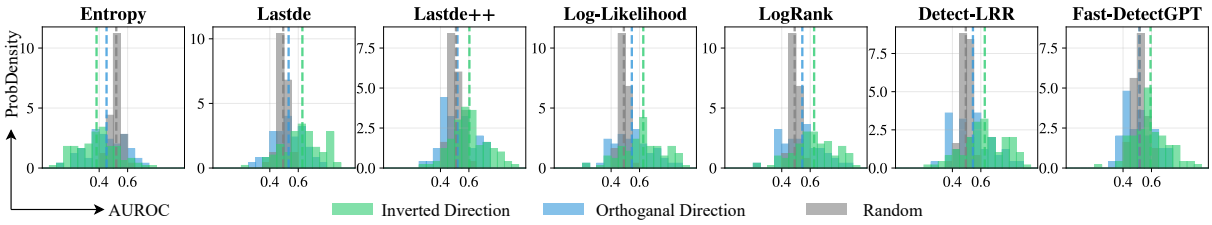

Figure 11: AUROC of each detector on three synthetic test sets and the distribution of feature value difference $\mathcal{D}(\cdot, \mathbf{w}^\star)$ under each construction.

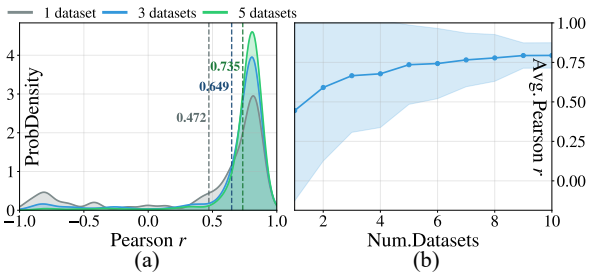

Figure 12: (a) Distribution of Pearson r with 1, 3, and 5 probe datasets. (b) Change of mean Pearson r with the number of probe datasets, shaded with one standard deviation intervals.

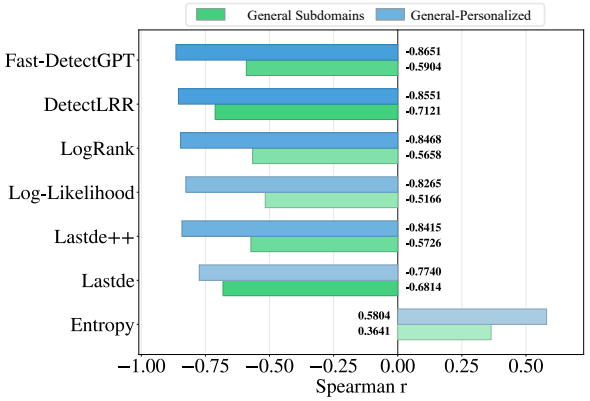

Figure 13: Correlation differences in different detectors.

that do not remain stable across domains. For example, detectors often assume that HWT is more diverse than MGT. Personalization alters this pattern and causes the same cue to change its meaning. This resembles spurious correlations, but the key difference is that the correlation does not simply weaken but is reversed across domains.

Domain Shift Robustness. The feature inversion phenomenon is also connected to domain shift. In most domain shift settings, model performance changes gradually because discriminative features lose reliability (Zhou et al., 2023). In personalization, the direction of the discriminative feature itself changes, and this leads to a reversal of predictions. This reflects a stronger and more structural shift than what is commonly observed in standard OOD cases, as also discussed in Appendix E.1.

Adversarial Robustness. For the feature-inversion trap, there is a conceptual relation to adversarial robustness because both involve changes that redirect model attention toward different features (Ilyas et al., 2019). Personalization modifies stylistic properties consistently, thereby altering how detectors use specific cues. However, the shift arises naturally from personalized text generation rather than from intentional adversarial manipulation, and the mechanisms are therefore different.

Invariant Representation Learning. Invariant representation learning aims to identify features that remain stable across domains. (Zhao et al., 2022) The feature inversion phenomenon shows that many detectors rely on features that are not invariant, including patterns related to lexical or semantic diversity. Even if invariant learning is applied, detectors may still depend on unstable cues when the degree of personalization is strong. This highlights a limitation of invariant feature learning in the context of MGT detection.

Feature Inversion in Natural Language Processing. The term feature inversion in natural language processing is commonly used to describe the reconstruction of input text from internal representations (Morris et al., 2023). Our use of the term is different. Here, "feature inversion" refers to a change in the discriminative direction of a feature across domains. A feature that separates HWT from MGT in one domain may work in the opposite way in another domain. This concerns a semantic reversal rather than the reconstruction of text.

E.3 What Do Inverted Features Capture

We believe that the inverted features include, but are not limited to, lexical diversity and semantic coherence. Both of these properties describe aspects of textual coherence, and both show a clear reversal trend across domains.

Lexical Diversity We follow the intrinsic dimension analysis from prior work (Tulchinskii et al., 2023), which reports that token embeddings of MGT usually lie on lower-dimensional manifolds than those of HWT. This indicates that MGT is generally less diverse in word usage. Using the probe datasets in Section 4.1, we follow the analysis procedure in (Tulchinskii et al., 2023). We use `Roberta-base` to encode each sentence. We take the hidden states from the last layer and treat the embedding at each token position as a point in the hidden space. We then apply persistent homology dimension (Schweinhart, 2021) to these points to estimate the dimensionality of lexical variation. Results are shown in Table 7. For HWT, personalized samples show higher lexical diversity than general ones. For MGT, diversity is lower in general settings but higher in personalized settings. This pattern is reversed across domains and aligns with the feature inversion phenomenon. It suggests that in highly personalized scenarios, MGT may use words in a more varied and aggressive manner than HWT.

Token Dimension	General	Personalized	Δ
HWT	9.9952	12.0255	2.0304
MGT	6.5751	13.0849	6.5098
Δ	-3.4201	1.0594	-

Table 7: Token-level intrinsic dimension for general and personalized domains.

Semantic Coherence We apply a similar idea to analyze sentence-level semantic coherence. With SentenceBERT (Reimers and Gurevych, 2019), each sentence can be mapped to a single vector. Closer vectors indicate stronger semantic consistency between sentences. If these vectors lie on a low-dimensional manifold, the text is more coherent, and the semantic flow is smoother. If the manifold has higher dimensionality, the text contains larger semantic shifts. We segment each text into sentences using punctuation and encode them with a SentenceBERT model[3], then estimate the manifold dimension for each group. Results are presented in Table 8. For MGT, coherence remains lower in general domains, but in personalized domains, it becomes higher than that of HWT. This further suggests that inverted features are related to both lexical diversity and semantic coherence.

[3]Here we choose the `all-MiniLM-L6-v2` model.

Sentence Dimension	General	Personalized	Δ
HWT	5.2764	4.4390	-0.8374
MGT	3.2783	5.4325	2.1542
Δ	-1.9981	0.9935	-

Table 8: Sentence-level intrinsic dimension for general and personalized domains.

E.4 Towards Mitigating the Feature Inversion Phenomenon

We suggest that finetuning on highly personalized text can improve detector accuracy within the target domain, but the improvement does not transfer across domains. Training on one domain does not yield gains in another, and joint training across domains cannot overcome the model's inherent limits in generalization. Thus, training alone cannot fully address the feature inversion phenomenon. An effective solution still requires either a method that directly captures inverted feature behavior or a detector that avoids relying on such unstable cues.

We evaluate three finetuning variants of Roberta. For `M4`, same as in Appendix B.3, each subset is randomly split by selecting 1,000 samples for testing, with all remaining samples used for training. `StyloBench` is split in the same way. The first variant is trained only on M4, the second only on `Stylo-Literary`, and the third on the combined training sets. All variants are then tested on the M4 test sets and on both parts of `StyloBench`. The results are shown in Table 9. Finetuning improves performance inside the training domain, but the gains do not transfer across domains. Even joint training on both domains does not close the gap, especially on `Stylo-Blog`. This shows that limited generalization ability continues to restrict the mitigation of the inversion phenomenon.

Training Set	Stylo-Literary	Stylo-Blog	M4
None	52.37	53.87	41.68
M4	53.33	64.77	99.94
Stylo-Literary	99.96	67.95	70.72
Mixed	99.95	71.94	99.95

Table 9: Performance of Roberta variants, trained on different data.

Detector \ Author	J.A	C.D	F.D	P.L	B.S	J.S	M.T	Avg.
Llama-3.1-8B								
Entropy	71.91	57.64	44.19	58.80	52.07	52.12	49.85	55.23
Lastde	15.17	72.96	86.64	72.80	85.08	73.41	83.10	69.88
Lastde++	12.85	58.34	76.84	58.45	76.40	66.63	73.17	60.38
Log-Likelihood	5.33	31.81	52.93	32.42	45.10	44.07	44.47	36.59
LogRank	6.07	34.62	54.38	34.35	47.56	45.67	47.08	38.53
Detect-LRR	13.23	46.21	57.74	41.72	55.74	50.79	54.67	45.73
Fast-DetectGPT	8.02	27.46	44.01	31.84	44.26	37.03	39.88	33.22
Phi-4								
Entropy	50.79	57.71	44.76	50.86	54.62	52.45	52.24	51.92
Lastde	25.83	65.61	78.99	66.94	78.14	66.27	77.95	65.67
Lastde++	13.40	42.14	61.89	43.13	60.90	51.73	59.78	47.57
Log-Likelihood	9.39	25.97	43.53	31.28	34.01	38.76	33.63	30.94
LogRank	11.41	28.60	44.26	32.22	35.73	39.98	36.02	32.60
Detect-LRR	26.85	40.34	47.65	36.90	43.69	44.38	44.56	40.62
Fast-DetectGPT	6.08	15.61	24.42	15.05	25.34	21.08	21.67	18.47
Qwen-3-4B								
Entropy	70.23	80.26	71.89	75.36	79.37	73.63	82.53	76.18
Lastde	23.59	60.30	72.78	68.88	76.44	61.58	74.43	62.57
Lastde++	9.88	30.06	46.55	42.82	54.65	41.56	52.94	39.78
Log-Likelihood	1.72	5.68	11.97	9.00	10.62	17.14	8.52	9.23
LogRank	2.42	7.15	12.44	9.90	12.23	18.68	10.26	10.44
Detect-LRR	11.67	17.30	19.39	17.96	23.16	26.64	19.88	19.43
Fast-DetectGPT	2.91	5.29	8.67	8.78	15.34	8.26	11.72	8.71

Table 10: Full AUROC results on `Stylo-Literary`, where generators are ***highlighted***.

Detector \ Subdomain	arXiv	PeerRead	Reddit	WikiHow	Wikipedia	Avg.
BLOOMz						
Entropy	33.95	23.66	23.82	41.90	83.33	41.33
Lastde	98.98	98.94	96.38	55.84	90.03	88.03
Lastde++	98.56	83.96	79.51	52.57	89.98	80.91
Log-Likelihood	88.71	68.97	64.56	50.05	26.44	59.75
LogRank	94.39	84.96	80.29	51.85	35.62	69.42
Detect-LRR	98.93	98.56	97.42	57.08	69.69	84.34
Fast-DetectGPT	96.04	46.73	31.23	25.41	75.66	55.01
ChatGPT						
Entropy	61.64	4.21	10.35	7.04	48.49	26.35
Lastde	99.00	99.00	99.37	92.62	97.40	97.48
Lastde++	99.85	99.85	99.72	94.86	99.07	98.67
Log-Likelihood	89.13	99.45	98.98	96.57	84.69	93.76
LogRank	91.48	99.51	99.14	96.26	88.29	94.94
Detect-LRR	94.08	99.29	98.96	93.02	95.32	96.13
Fast-DetectGPT	99.93	99.92	99.84	96.22	99.03	98.99
Cohere						
Entropy	64.16	14.39	28.84	6.62	45.16	31.83
Lastde	95.91	97.28	99.67	99.57	96.01	97.69
Lastde++	98.50	95.93	98.74	99.90	98.04	98.22
Log-Likelihood	85.67	97.02	96.08	99.43	87.39	93.12
LogRank	88.24	96.92	96.38	99.56	89.97	94.21
Detect-LRR	90.59	95.28	96.13	99.37	93.04	94.88
Fast-DetectGPT	99.50	97.00	99.20	99.93	98.30	98.78
Davinci						
Entropy	78.65	19.92	17.00	23.26	61.68	40.10
Lastde	50.53	99.49	94.32	85.46	88.73	83.70
Lastde++	41.47	99.81	98.34	87.80	94.64	84.41
Log-Likelihood	21.97	99.11	95.22	84.26	63.40	72.79
LogRank	20.68	99.30	95.39	83.46	67.08	73.18
Detect-LRR	20.29	99.25	93.77	78.75	78.10	74.03
Fast-DetectGPT	44.09	99.91	98.72	89.36	94.29	85.28

Table 11: Full AUROC results on `M4`, where generators are ***highlighted***.

Detector	Generator	J.A	C.D	F.D	P.L	B.S	J.S	M.T	Avg.
Roberta	Llama-3.1-8B	35.09	66.87	68.81	46.90	64.12	50.71	63.81	56.62
	Phi-4	46.23	55.77	61.65	36.71	55.15	45.10	54.50	50.73
	Qwen3-4B	58.82	58.25	54.92	37.25	57.25	45.75	56.27	52.64
DeTeCtive	Llama-3.1-8B	68.98	65.31	61.80	66.34	71.73	64.02	63.21	65.91
	Phi-4	63.63	63.09	68.25	65.75	69.15	63.40	65.10	65.48
	Qwen3-4B	82.99	77.16	77.41	71.48	75.94	74.01	70.29	75.61

Table 12: Results of training-based detectors on `Stylo-Literary`.

Detector	Claude3.7	Claude4	GPT-4o	Qwen2.5-72B	Avg.
Roberta	47.38	57.33	59.40	94.99	64.78
DeTeCtive	43.97	45.06	47.68	60.39	49.28

Table 13: Results of training-based detectors on `Stylo-Blog`.

Detector	Generator	arXiv	PeerRead	Reddit	WikiHow	Wikipedia	Avg.
Roberta	Davinci	99.33	100.00	100.00	100.00	99.98	99.86
	Cohere	99.92	99.99	99.95	100.00	99.77	99.93
	ChatGPT	100.00	100.00	100.00	100.00	99.95	99.99
	BLOOMZ	99.98	100.00	100.00	100.00	100.00	100.00
DeTeCtive	Davinci	21.04	99.56	71.45	38.15	49.12	55.86
	Cohere	67.04	98.64	86.20	76.87	40.97	73.94
	ChatGPT	89.73	99.83	80.51	39.62	56.75	73.29
	BLOOMZ	70.95	96.80	87.49	49.23	91.62	79.22

Table 14: Results of training-based detectors on the English subset of M4.

Table 15: Examples of `Stylo-Blog`. The first 30 tokens used to generate MGTs are **bolded**.

Model	Text
	Human
Human	There is so much on my mind these days... So much that I can't seem to form a clear thought. I want so badly to move on with my life, but habits and people and I, myself, keep pulling me back. I know what I need to do, but it's almost like I am afraid to take that step. I've always done the safe thing, made the safe choices. I've always done what was expected of me. And now that I am finally trying to step out on my own, I'm being told that I am running away. As much as I know that I need to do this, I still have those same childish fears that I've always had. I love my family, I love my friends, but I use them as a crutch. I use them as the reasons to stay in my comfort zone. I use them as my excuses to not step into that shadowy place just beyond my line of sight. I am not running away from anything, well, maybe some things, but sometimes, running away is the only way to set yourself free. I've prayed and prayed, begged and pleaded with God to not tell me what I already know. There is no way to deny that I know where my life is going. The only way that I can get to where I truly need to be is to walk away from everything that I know, to take that first blind step into my future. And the only way to do that is to put my complete trust and total faith in God.
	Machine

Continued on next page

Table 15: Examples of `Stylo-Blog` (continued)

Model	Text
Qwen2.5-72B	**There is so much on my mind these days... So much that I can't seem to form a clear thought. I want so badly to** share all these swirling emotions and thoughts, but the words just don't seem to come. It feels like a storm is brewing inside me, with no outlet in sight. The more I try to wrap my head around everything, the more I feel like I'm losing my grip on the present. \n\nI've been thinking a lot about the people in my life recently. The relationships that have shaped me, the ones that have endured through thick and thin, and the ones that have faded into the background. It's funny, how the people who are closest to you can be the ones you struggle to open up to the most. I find myself holding back, afraid of what will happen if I let it all out. But then, there are moments, like that one in church, where everything bubbles up to the surface and I can't hold it back.\n\nLast Sunday, I was standing there, singing those familiar worship songs, and it was like I was transported. The words, \"I walk by faith, each step by faith; to live by faith, I put my trust in You,\" echoed in my heart. And in that moment, I felt it. I looked around, and the faces of the people surrounding me were both familiar and strange. Familiar, because they've been a part of my life for so long. Strange, because life has a way of changing us, sometimes so subtly that we don't even notice until we stop and really look.\n\nOf the couple hundred people in that room, I only knew about half. Of those, there were the people who have been constants in my life: the adults who have watched me grow from a girl into a woman, and the children who I have watched grow from toddlers to young adults. There were the people who have stood by me through my weakest moments, holding me up when I thought I would fall. And there were those who I have supported, giving them a bit of my strength when they needed it most.\n\nBut there was one person, one person who I couldn't talk to, the one person I wasn't sure if I could leave. And in that moment, the tears started to sting my eyes. I couldn't reach out, couldn't communicate what I was feeling. The song ended in fellowship time, and I
Claude-4	**There is so much on my mind these days... So much that I can't seem to form a clear thought. I want so badly to** move on with my life, but I feel like I'm stuck in this weird limbo where nothing makes sense anymore. You know that feeling when you're standing at a crossroads and every direction looks both terrifying and exciting at the same time? That's exactly where I am right now.\n\nI keep thinking back to all those times when I was so sure I had everything figured out. Remember when I was eighteen and thought I knew exactly how my life would unfold? Ha! If only that naive girl could see me now. She'd probably be shocked at how different things turned out, but maybe she'd also be proud of the woman I've become through all the unexpected twists and turns.\n\nThe truth is, I'm tired of feeling like I'm waiting for permission to live my own life. Does that make sense? It's like I've been holding my breath, waiting for some magical moment when everything will click into place and I'll suddenly know what I'm supposed to do next. But what if that moment never comes? What if this uncertainty is just part of being human, part of growing up, part of learning to trust yourself even when you can't see the whole staircase?\n\nI was talking to my mom the other day, and she said something that really stuck with me. She said, \"Honey, sometimes moving on doesn't mean you have all the answers. Sometimes it just means you're brave enough to take the next step without
Claude-3.7	**There is so much on my mind these days... So much that I can't seem to form a clear thought. I want so badly to** just sit and process everything, but life keeps moving at this relentless pace that doesn't allow for much reflection.\n\nYou know that feeling when your brain is like a browser with too many tabs open? That's me right now. A hundred thoughts racing, memories surfacing at odd moments, worries about tomorrow pressing in, and in the midst of it all, trying to remember if I picked up milk at the store yesterday.\n\nI found myself driving home from work yesterday, and I couldn't even tell you what streets I took. My body was on autopilot while my mind wandered through a maze of its own making. It's scary sometimes, how disconnected we can become from the present moment.\n\nLast night, I was folding laundry – such a mundane task – when suddenly tears were streaming down my face. Not sad tears necessarily, just... release tears. The kind that come when your heart is too full of something unnamed. The kind that need no explanation or justification. They just are.\n\nAnd in that moment, with a half-folded t-shirt in my hands, I realized how much I've been holding in. How many emotions I've been packing away neatly like these clothes, trying to make everything fit in the limited space I've allowed myself.\n\nI think we all do this sometimes. We keep pushing forward, keep smiling, keep showing up for the people who need us. We sing the worship songs on Sunday morning, we make the lunches, we answer the emails, we check the boxes. And somewhere in all of that, our own hearts get a little lost.\n\nI'm looking at my Bible sitting on the nightstand. It's been there for days, untouched. Not because I don't want to open it, but because I'm afraid of what might happen if I do. If I get still enough to listen, what will God say? What will He ask of me? What emotions might surface that I've been working so hard to keep submerged?\n\nI walk by faith, each step by faith.\n\nThose words from the song at church keep coming back to me. Walking by faith doesn't always mean having clear direction. Sometimes it means putting one foot in front of the other when the path ahead is completely obscured. Sometimes it means trusting that even in the mental chaos, God

Continued on next page

Table 15: Examples of `Stylo-Blog` (continued)

Model	Text
GPT-4o	**There is so much on my mind these days... So much that I can't seem to form a clear thought. I want so badly to** put it all into words, but every time I try, they swirl around like leaves caught in an autumn breeze. It's as if the very alphabet refuses to settle on the page, preferring instead to dance in chaos. And amidst this whirlwind, I find myself yearning for clarity, for a moment of stillness where I can breathe and just... be.\n\nSometimes I think about how life used to be so simple, yet we couldn't wait to grow up. Oh, how naive we were! The dreams of adulthood seemed wrapped in wonder and possibility, but now I see that they come twined with responsibility and introspection. I pause, caught in the space between what was and what is, longing for that childlike clarity once more.\n\nThis introspection takes me back to those quiet moments in the pews of the old church, where the weight of expectations lifted slightly with each hymn sung. There, in the company of familiar faces, I found something akin to peace. We were together, bound not just by belief but by the shared journey of faith and vice-versa, bearing the burdens of our collective humanity.\n\nYet these days, everywhere I turn, there seems to be complexity and noise. The world spins faster, and sometimes I feel like an outsider peering in through fogged glass. I find solace in small acts: the laughter of a child, the rustle of leaves, the comforting ordinariness of a daily routine. I seek meaning in the minutiae, an anchor in the intricate tapestry of life.\n\nI've been volunteering more with the local youth group, hoping to give back some of what was freely given to me. And, oh, how their perspectives refresh my own! Their unfiltered view of the world serves as a reminder of what truly matters. I watch them, these bundles of potential and enthusiasm, and think about possibilities — not just their possibilities, but mine too. \n\nHow often do we, as adults, lose sight of our dreams? They get locked away like treasures in a dusty attic, while we pursue notions of success that may not really belong to us. And yet, every now and then, I'm reminded to dust them off, to recall who I truly am beyond the roles and titles accumulated over the years. It's like catching a glimpse of your reflection in

Table 16: Examples of `Stylo-Literary`. The first 30 tokens used to generate MGTs are **bolded**.

Source	Text
	Human Author: Mark Twain Generator: Llama-3.1-8B
Human	**"Say, wouldn't it 'a' been nuts if he'd a-been here last night " "Oh, by George, but wouldn't** it" said Ferguson "Then we'd have seen scientific work Intellect–just pure intellect–away up on the upper levels, dontchuknow Archy is all right, and it don't become anybody to belittle him, I can tell you But his gift is only just eyesight, sharp as an owl's, as near as I can make it out just a grand natural animal talent, no more, no less, and prime as far as it goes, but no intellect in it, and for awfulness and marvelousness no more to be compared to what this man does than–than–Why, let me tell you what he'd have done He'd have stepped over to Hogan's and glanced–just glanced, that's all–at the premises, and that's enough See everything Yes, sir, to the last little detail; and he'll know more about that place than the Hogans would know in seven years Next, he would sit down on the bunk, just as ca'm, and say to Mrs Hogan–Say, Ham, consider that you are Mrs Hogan I'll ask the questions; you answer them " "All right; go on " "'Madam, if you please–attention–do not let your mind wander Now, then–sex of the child ' "'Female, your Honor ' "'Um–female Very good, very good Age ' "'Turned six, your Honor ' "'Um–young, weak–two miles Weariness will overtake it then It will sink down and sleep We shall find it two miles away, or less Teeth ' "'Five, your Honor, and one a-coming ' "'Very good, very good, very good, indeed ' You see, boys, he knows a clue when he sees it, when it wouldn't mean a dern thing to anybody else 'Stockings, madam Shoes ' "'Yes, your Honor–both ' "'Yarn, perhaps Morocco ' "'Yarn, your Honor And kip ' "'Um–kip This complicates the matter However, let it go–we shall manage Religion ' "'Catholic, your Honor ' "'Very good Snip me a bit from the bed blanket, please Ah, thanks
Machine	**"Say, wouldn't it 'a' been nuts if he'd a-been here last night " "Oh, by George, but wouldn't** that be too many, though I like to talk it out once, just for pastime; and sometimes, when I'm alone in the shop, I try to make it all come again in dialog, and then it seems as if it had, and I dream about it nights You can imagine yourn over again, same as I do When I get tired of myn, I take up Huck's " "Why, it was the most–" "Yes–incoherency, that's what it was They just rambled around from one thing to another, and never landed in the same place twice Let it alone; it ain't anybody's fault Consider them illiterates Now you shell notice, Astronomical would naturally drop into their own language; and he did Then he would always try to correct himself, and begin to sling off parts of words that he thought belonged to English, but didn't No, he couldn't put together so you could remember anything he'd said, within fifteen minutes after it was done; and he couldn't get his sentence started twice, right straight through So somebody else would interrupt and say, "What's de use, let him 'lone–dey can't git shet o' de subject," or something like that, you know, and then start himself and go to talking You see, he tried to talk English all the time, only he couldn't somehow It got him down So he took to talking dialect, because it was easier and convenient, you know Well, as soon as he'd break loose, there warn't no getting him to his english again He couldn't seem to settle down to it We all used to go for him and rush him every time he'd begin to talk English, and force him to talk dialect The Widow Douglas used to break in on me dreadful at first, with her English; but she got the hang of it a deal sooner than any of the rest She said she couldn't stand it, and she warned me off from trying to make a body talk dialect that hadn't been brought up to it, I might as well atrophy an elephant Yes, even the widow: everybody was down to the other fellow's level, before we got through And some better'n others, too Present
	Human Author: Jane Austen Generator: Llama-3.1-8B
Human	**" So far her improvement was sufficient–and in many other points she came on exceedingly well; for though she could not write sonnets, she brought herself to read them**; and though there seemed no chance of her throwing a whole party into raptures by a prelude on the pianoforte, of her own composition, she could listen to other people's performance with very little fatigue Her greatest deficiency was in the pencil–she had no notion of drawing–not enough even to attempt a sketch of her lover's profile, that she might be detected in the design There she fell miserably short of the true heroic height At present she did not know her own poverty, for she had no lover to portray She had reached the age of seventeen, without having seen one amiable youth who could call forth her sensibility, without having inspired one real passion, and without having excited even any admiration but what was very moderate and very transient This was strange indeed But strange things may be generally accounted for if their cause be fairly searched out There was not one lord in the neighbourhood; no–not even a baronet There was not one family among their acquaintance who had reared and supported a boy accidentally found at their door–not one young man whose origin was unknown Her father had no ward, and the squire of the parish no children But when a young lady is to be a heroine, the perverseness of forty surrounding families cannot prevent her Something must and will happen to throw a hero in her way Mr Allen, who owned the chief of the property about Fullerton, the village in Wiltshire where the Morlands lived, was ordered to Bath for the benefit of a gouty constitution–and his lady, a good-humoured woman, fond of Miss Morland, and probably aware that if adventures will not befall a young lady in her own village, she must seek them abroad, invited her to go with them Mr and Mrs Morland were all compliance, and Catherine all happiness

Examples of `Stylo-Literary`, continued from last page.

Source	Text
Machine	**" So far her improvement was sufficient–and in many other points she came on exceedingly well; for though she could not write sonnets, she brought herself to read them** with feeling, and the history of her first setting out to excel was soon written, and made into a narrow little volume, two volumes in quinto, and all in her own hand She had no notice of it herself, till her mother one day said, "Ellen, may I see your drawing-room window " And when Elizabeth turned her hand out of the draw- ing-book, some elegant little figure of a tower or grotto which Miss Crawford had criticising, there, without thinking, lay the pretty little volume The look which Elizabeth instantly raised to comprehend the business, remained some time on Miss Crawford's face, and hers was AN UNPLEASANT ONE It spoke meekness, submitting, shy distrust, penetrating curiosity, but yet grudging invitation A second or two passed in complete silence before either of them spoke, and then it was only, "Well, Ellen –I thought you had been writing " "Oh I did, indeed, I was so much concerned about it " "And have you explained everything to Mr Morland What does he say " "He is very good-aided But I have been used to have poorer friends My family have been poor, for years " "Yes; that is, you have no father or brother living I beg your pardon You are on the spot yourself, which is always something in a friend He must be a singular man–a silent, melancholy, earnest, persevering, thinking, practicing man–his mind completely under his own control–somewhat severe, perhaps–but with great energy–great forbearance–great capacity of enjoying the thing that he likes –provided a woman's wishes were consulted but what do I tell you They never ARE At this rate you will never know Badly as we judge alone, we should never dare to judge till
Human Author: Jane Austen Generator: Phi-4	
Human	**" "I am very sorry for it," said Catherine dejectedly; "if I had known this before–" "Perhaps," said Miss Tilney** in an embarrassed manner, "you would be so good–it would make me very happy if–" The entrance of her father put a stop to the civility, which Catherine was beginning to hope might introduce a desire of their corresponding After addressing her with his usual politeness, he turned to his daughter and said, "Well, Eleanor, may I congratulate you on being successful in your application to your fair friend " "I was just beginning to make the request, sir, as you came in " "Well, proceed by all means I know how much your heart is in it My daughter, Miss Morland," he continued, without leaving his daughter time to speak, "has been forming a very bold wish We leave Bath, as she has perhaps told you, on Saturday se'nnight A letter from my steward tells me that my presence is wanted at home; and being disappointed in my hope of seeing the Marquis of Longtown and General Courteney here, some of my very old friends, there is nothing to detain me longer in Bath And could we carry our selfish point with you, we should leave it without a single regret Can you, in short, be prevailed on to quit this scene of public triumph and oblige your friend Eleanor with your company in Gloucestershire I am almost ashamed to make the request, though its presumption would certainly appear greater to every creature in Bath than yourself Modesty such as yours–but not for the world would I pain it by open praise If you can be induced to honour us with a visit, you will make us happy beyond expression 'Tis true, we can offer you nothing like the gaieties of this lively place; we can tempt you neither by amusement nor splendour, for our mode of living, as you see, is plain and unpretending; yet no endeavours shall be wanting on our side to make Northanger Abbey not wholly disagreeable " Northanger Abbey These were thrilling words, and wound up Catherine's feelings to the highest point of ecstasy Her grateful and gratified heart could hardly restrain its expressions within the language of tolerable calmness To receive so flattering an invitation To have her company so warmly solicited
Machine	**" "I am very sorry for it," said Catherine dejectedly; "if I had known this before–" "Perhaps," said Miss Tilney**, with an affectionate smile, "it would not have made you more easy on the subject " Her brother instantly interfered; "Do you advise her to tell me of it then " "By no means, if she can help it The disadvantage is at least mutual There need be no consequence but what will be attended by pleasure " But though the matter was apparently arranged in a moment, and they all three seemed most happily reconciled to each other, there yet lingered in Catherine's mind some feelings which required the subduing influence of a good night's rest to put them quite to flight; and, though without exactly knowing why, as soon as she left the room, and while slowly ascending the staircase, she felt convinced that General Tilney could never like her, or think her pretty –They had one family friend, besides these gentlemen, who resided among them Mr Frederick Tilney, their eldest son, was the rector of the parish, of which his father was patron A young man very like his cousin John, but handsomer than him He had a great deal of the same manner–and every body thought him extremely agreeable, except Catherine She did not like him; and, now that Edmund Bertram was removed from her acquaintance, she wished very much that _he_ might never return from Oxford It appeared afterwards that he must often visit his relations in Gloucestershire, and this raised a thousand unpleasant questions: Whom would he come with Why was not his sister invited also Would he bring Anne Heretitia Mention of Mr Allen did so little honour to either sex, to human nature in general, that Catherine was rather glad the circumstance was suppressed They sat down to table, twenty-three in number, and in a style of luxury, which, to Catherine, with the exception of her first morning's entertainment, was peculiar Having been indulged with uncommon orders in her favourite toy-shop only two days before, she knew how much superior was cold beef, or cold lamb, at any time of life, to the daintiest dinner that ever was cooked; and being prepared, therefore, to find fault, nothing could be more exquisite than her admiration In addition to every other

Examples of `Stylo-Literary`, continued from last page.

Source	Text
	Human Author: Bernard Shaw Generator: Phi-4
Human	**We shall come to the question of the economic dependence of women on men later on; but at present we had better confine ourselves to the theories of** marriage which we are not ashamed to acknowledge and defend, and upon which, therefore, marriage reformers will be obliged to proceed We may, I think, dismiss from the field of practical politics the extreme sacerdotal view of marriage as a sacred and indissoluble covenant, because though reinforced by unhappy marriages as all fanaticisms are reinforced by human sacrifices, it has been reduced to a private and socially inoperative eccentricity by the introduction of civil marriage and divorce Theoretically, our civilly married couples are to a Catholic as unmarried couples are: that is, they are living in open sin Practically, civilly married couples are received in society, by Catholics and everyone else, precisely as sacramentally married couples are; and so are people who have divorced their wives or husbands and married again And yet marriage is enforced by public opinion with such ferocity that the least suggestion of laxity in its support is fatal to even the highest and strongest reputations, although laxity of conduct is winked at with grinning indulgence; so that we find the austere Shelley denounced as a fiend in human form, whilst Nelson, who openly left his wife and formed a menage a trois with Sir William and Lady Hamilton, was idolized Shelley might have had an illegitimate child in every county in England if he had done so frankly as a sinner His unpardonable offence was that he attacked marriage as an institution We feel a strange anguish of terror and hatred against him, as against one who threatens us with a mortal injury What is the element in his proposals that produces this effect The answer of the specialists is the one already alluded to: that the attack on marriage is an attack on property; so that Shelley was something more hateful to a husband than a horse thief: to wit, a wife thief, and something more hateful to a wife than a burglar: namely, one who would steal her husband's house from over her head, and leave her destitute and nameless on the streets
Machine	**We shall come to the question of the economic dependence of women on men later on; but at present we had better confine ourselves to the theories of** Mrs Warren's profession As far as its religious aspect goes, it seems, so far as I can make out, to be a practice of prostitution in which no payment is accepted and from which the practitioner expects nothing but damnation This has never been a very popular method: St Paul tells us that prostitutes were always ready enough to trade for money without asking for any recompense in heaven Consequently it was not until Mr Gladstone found out, when he visited Florence in 1897, that Fra Angelico represented angels with wings instead of tails, that a reaction set in against Dante, who did not take this view of things There are many saints in our calendar, male and female, whose sanctity has apparently depended largely upon their experience of this form of the calling But these records are extremely obscure, because they all insist much more strongly than Dante does on the fact that there is only one thing wrong with vice: that is, that it should pay Now in all discussions as to whether there is anything fundamentally wrong with human nature–the strongest practical grounds being generally alleged by those who believe nothing whatever in its fundamental soundness–it is necessary to have a common denominator Both sides must agree what ẅrongm̈eans before they can go on to argue about the degree in which it exists If there is something really wrong with human nature, then, logically, no human conduct or instinct or desire can possibly be regarded as infallibly right The immorality of an action cannot depend entirely on the motive behind it You may begin by admitting that if you do certain things with a good motive your act is not sinful; but still, you cannot admit that such acts done with a bad motive are not wicked Why should not the most disinterested and self-sacrificing promiscuity be just as vicious as the greediest and vilest lust And why should not the vilest and greediest lust be just as virtuous as the purest devotion Can anyone answer these questions sensibly (I am sure nobody ever tries) except on the
	Human Author: Bernard Shaw Generator: Qwen3-4B

Examples of `Stylo-Literary`, continued from last page.

Source	Text
Human	**We shall come to the question of the economic dependence of women on men later on; but at present we had better confine ourselves to the theories of** marriage which we are not ashamed to acknowledge and defend, and upon which, therefore, marriage reformers will be obliged to proceed We may, I think, dismiss from the field of practical politics the extreme sacerdotal view of marriage as a sacred and indissoluble covenant, because though reinforced by unhappy marriages as all fanaticisms are reinforced by human sacrifices, it has been reduced to a private and socially inoperative eccentricity by the introduction of civil marriage and divorce Theoretically, our civilly married couples are to a Catholic as unmarried couples are: that is, they are living in open sin Practically, civilly married couples are received in society, by Catholics and everyone else, precisely as sacramentally married couples are; and so are people who have divorced their wives or husbands and married again And yet marriage is enforced by public opinion with such ferocity that the least suggestion of laxity in its support is fatal to even the highest and strongest reputations, although laxity of conduct is winked at with grinning indulgence; so that we find the austere Shelley denounced as a fiend in human form, whilst Nelson, who openly left his wife and formed a menage a trois with Sir William and Lady Hamilton, was idolized Shelley might have had an illegitimate child in every county in England if he had done so frankly as a sinner His unpardonable offence was that he attacked marriage as an institution We feel a strange anguish of terror and hatred against him, as against one who threatens us with a mortal injury What is the element in his proposals that produces this effect The answer of the specialists is the one already alluded to: that the attack on marriage is an attack on property; so that Shelley was something more hateful to a husband than a horse thief: to wit, a wife thief, and something more hateful to a wife than a burglar: namely, one who would steal her husband's house from over her head, and leave her destitute and nameless on the streets
Machine	**We shall come to the question of the economic dependence of women on men later on; but at present we had better confine ourselves to the theories of** immorality and bad manners which is, in short, one single doctrine which has grown by multiplication into an infinity of articles To all these it may be said that the theory to which they are attributable is the theory of physical inferiority, moral incapacity, inhumanity or mere unskilled destructiveness The elementary creed includes not only the theory of male superiority as actually evidenced in modern warfare, but also such absurd assumptions as that in spite of modern civilized warfare women cannot fight and must therefore remain separate from mankind even if they have no more powerful reason for doing so; or again that the splendid courage displayed by the women who stood out during the Great War was produced by a preternatural instinctive force: in other words, a sexual instinctuality, which is declared to be incapable of being cultivated by education All this is gobbledygook made up by professional female critics to please their paymasters The truth in short is what Adam Smith, long ago put it thus T̈o man's nature is due his love of dominion over woman He may, and I believe he ought sometimes to, allow her freedom of choice in marriage matters; but when he does this, he leaves his wife in much the same position as if she were married to him against her will A great deal of what we call justice or respectability between husband and wife really comes from a strong sense of self-reverence, accompanied by an equally strong sense of the infamy of a lie ẗHE HOMO SAPIENS MUSICAL PARADISE Some people feel perfectly happy in the belief that women are more refined than men because they are less brutal, less selfish, less rapacious In itself this idea makes some sort of sense; but it has been found that it does not explain everything Every human being desires, sooner or later, some sort of gratification Men usually seek satisfaction from sex activity, and sometimes from sublimated forms of that activity (theology, science, music); whereas women often find no satisfactory substitute except bare animal