Quantifying Social Biases in Language Model Classifiers is Domain-Dependent

TAMARA QUIROGA, Department of Computer Sciences, Pontifical Catholic University of Chile, Chile, Department of Computer Sciences, University of Chile, Chile, National Center for Artificial Intelligence, Chile, and Millennium Institute for Foundational Research on Data, Chile
FELIPE BRAVO-MARQUEZ, Department of Computer Sciences, University of Chile, Chile, National Center for Artificial Intelligence, Chile, and Millennium Institute for Foundational Research on Data, Chile
VALENTIN BARRIERE, Department of Computer Sciences, University of Chile, Chile and National Center for Artificial Intelligence, Chile

As natural language processing (NLP) systems increasingly operate on heterogeneous data, faithful social bias evaluation requires benchmarks that are both domain-sensitive and scalable. However, existing resources still face a core trade-off: template-based benchmarks scale easily but often lack linguistic authenticity, whereas naturally occurring examples (NOEs) reflect real usage but are expensive to curate and often provide limited coverage across domains and social groups. To bridge this gap, we investigate whether large language models (LLMs) can automatically adapt template-based bias datasets to specific domains using zero-shot prompting.

We evaluate the effectiveness of LLMs along three dimensions: bias consistency, which measures agreement between bias estimates derived from adapted templates and NOEs; content preservation, which assesses whether the original semantic intent is maintained; and domain adherence, which evaluates how well the generated text reflects the linguistic characteristics of the target domain. Across the Equity Evaluation Corpus (EEC) and the Identity Phrase Templates Test Set (IPTTS), and across three domains—IMDb, Twitter, and Wikipedia Talk Pages—using both automatic and human evaluations, we show that domain-adapted templates capture real-world bias patterns more faithfully than standard templates. Overall, our results highlight the importance of domain variation in fairness research and position LLM-based adaptation as a scalable and reproducible framework for robust social bias assessment in NLP.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Natural language generation**.

Additional Key Words and Phrases: fairness in NLP, social bias, large language models, template-based benchmarks, domain adaptation, style transfer

ACM Reference Format:
Tamara Quiroga, Felipe Bravo-Marquez, and Valentin Barriere. 2026. Quantifying Social Biases in Language Model Classifiers is Domain-Dependent. *ACM Trans. Intell. Syst. Technol.* 37, 4, Article 111 (August 2026), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' Contact Information: Tamara Quiroga, t.quiroga@uc.cl, Department of Computer Sciences, Pontifical Catholic University of Chile, Chile and Department of Computer Sciences, University of Chile, Chile and National Center for Artificial Intelligence, Chile and Millennium Institute for Foundational Research on Data, Chile; Felipe Bravo-Marquez, fbravo@ddc.uchile.cl, Department of Computer Sciences, University of Chile, Chile and National Center for Artificial Intelligence, Chile and Millennium Institute for Foundational Research on Data, Chile; Valentin Barriere, vbarriere@ddc.uchile.cl, Department of Computer Sciences, University of Chile, Chile and National Center for Artificial Intelligence, Chile.

ACM 2157-6912/2026/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Social bias has been widely studied in NLP, leading to a substantial body of work on how to measure bias and compare mitigation strategies. These resources span a wide range of social dimensions [18, 24, 64, 65, 67] and task types [14, 43, 51]. Despite this progress, the role of domain in large-scale bias evaluation remains underexplored. Following Plank [52], we define a domain as a coherent corpus type characterized by systematic variation in topic, style, or register. In practice, NLP systems are deployed across diverse domains, ranging from the informal brevity of social media posts to the specialized terminology and formal conventions of scientific writing.

Existing bias benchmarks generally fall into two broad categories: template-based datasets and NOEs. Template-based datasets rely on fixed sentence patterns (e.g., "`The [profession] is [adjective]`") and are widely used because they are scalable and can cover many demographic groups and languages [7, 61, 62, 71]. However, they are often lexically constrained and may be distributionally mismatched with the domains in which models are ultimately applied [36, 60]. In contrast, NOEs are drawn from real-world corpora such as Wikipedia or Reddit and therefore better capture domain-specific language patterns [4, 14, 49, 70]. Yet they can be costly to construct and may be unevenly distributed across domains and demographic groups. As a result, achieving bias evaluation that is both domain-aware and scalable remains an open challenge.

Failing to account for domain in bias evaluation can distort the resulting estimates. Prior work on domain shift has shown that model behavior may change when evaluated on data drawn from different distributions [25, 33]. As a result, using datasets that are distant from the target domain can lead to misleading bias measurements and, consequently, to downstream risks when such estimates are used to guide real-world decisions. For example, this may affect content moderation, ranking and recommendation outcomes, and semantic decisions in high-stakes pipelines. Likewise, mitigation strategies may appear effective without adequately reflecting model behavior in real-world settings. At the same time, the absence of adaptable datasets limits progress toward a generalizable fairness assessment that can address diverse demographic groups and languages.

To address these limitations, we use LLMs to adapt scalable template-based bias datasets to specific domains through zero-shot prompting. Inspired by style transfer, our approach preserves the controllability and systematic coverage of templates while making them better reflect the linguistic characteristics of target domains.

To assess whether LLMs are effective for this purpose, we propose an evaluation framework based on three dimensions. Bias consistency measures whether bias estimates from real-world data in a target domain $\mathcal{D}$ are better approximated by the adapted template set $\mathcal{D}_{T_{\mathcal{D}}}$ than by the original template set $\mathcal{D}_T$ (Figure 1). The other two dimensions are content preservation, which evaluates whether the semantic meaning of the original templates is retained, and domain adherence, which captures the extent to which generated instances reflect the style of the target domain. For the latter two dimensions, we combine automatic metrics with human evaluation.

We apply our method to three domains: Twitter, Wikipedia Talk Pages (WTP), and IMDb movie reviews. We selected these domains for two main reasons. First, they contain frequent interpersonal and opinionated language, making them relevant settings where stereotypes and social biases may emerge [63, 69]. At the same time, they exhibit clear and systematic differences in length, orthography, and discourse structure.

Second, Twitter and WTP have been used in prior fairness research through template-based benchmarks that are not explicitly aligned with these domains [11, 15, 43, 50, 74]. This makes them particularly well suited for comparing bias estimates derived from domain-adapted benchmarks with those obtained from domain-agnostic ones.

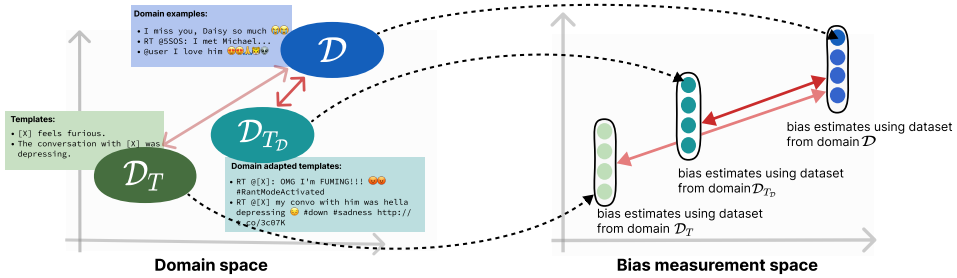


Fig. 1. Domain-adaptive framework for fairness evaluation. Templates from a source domain ($\mathcal{D}_T$) are adapted to a target domain ($\mathcal{D}$) using LLMs, producing domain-adapted templates ($\mathcal{D}_{T_\mathcal{D}}$). From these templates, bias vectors are derived—representations of disparities across protected groups—which are then compared against vectors obtained from real-world in-domain data ($\mathcal{D}$). This allows bias measurements that are sensitive to domain variation.

We apply our framework to the generation of templates for bias evaluation in settings where demographic groups are represented through names. Specifically, we focus on nationality and the intersection of gender and race/ethnicity, by adapting two widely used bias detection datasets for sentiment and hate speech models: EEC [30] and IPTTS [15].

Our evaluations show that, across domains, adapted templates produce bias estimates that are more consistent with NOEs than those obtained from the original templates. IPTTS adaptations generally show stronger alignment with NOEs, alongside better content preservation and domain adherence. We also find that adaptation yields larger improvements in domains with more distinctive registers and discourse structures. Building on our preliminary study [55], this work substantially broadens the empirical evaluation and provides a more comprehensive analysis of the proposed framework. Our contributions are threefold:

(1) We show across three domains that bias estimates are highly domain-sensitive and that domain-agnostic benchmarks can misrepresent real-world fairness risks.
(2) We evaluate LLM-based domain adaptation as a scalable approach for improving the realism of template-based benchmarks while preserving their controllability, and validate it along three dimensions: bias consistency, content preservation, and domain adherence.
(3) We assess the robustness of the framework across generator LLMs, downstream models, prompt configurations, and numbers of in-domain examples.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces our method and validation protocol, and Section 4 describes the experimental setup. Section 5 presents the main findings and discusses their implications, while Section 6 offers a deeper analysis of the method in terms of content preservation and domain transfer. Section 7 provides a geometric analysis of the approach. Finally, Section 8 concludes the paper and outlines directions for future work. The code is publicly available [1].

2 Related Work

Our work builds on prior research in social biases in NLP, including datasets for systematic bias measurement and the use of names as demographic proxies. It also draws on advances in style transfer with LLMs, which enable adapting text to domain-specific linguistic contexts.

[1] https://github.com/TamaraQuiroga/AdaptedTemplates

2.1 Social Biases in NLP

2.1.1 Measuring Social Biases. A wide range of datasets have been developed to assess social biases in NLP systems [19, 20, 24, 38, 65]. Many of these rely on counterfactual inputs—pairs or sets of sentences with equivalent meanings that differ only in the social groups they reference. Bias is then inferred from differences in model outputs across these groups. Broadly, such datasets are constructed using three main strategies: (i) template-based, (ii) naturally occurring examples, and (iii) crowdsourcing.

Template-based. Built from manually crafted sentences, template-based datasets are widely used due to their scalability and controllability. They have facilitated research on underexplored dimensions of bias, such as nationality and disability [43, 62], as well as extensions to languages beyond English [12, 29]. However, their synthetic nature makes them limited: templates are lexically constrained, often out-of-distribution [6, 36, 60], and also have been criticized for lack of reliability [59, 60] and weak correlation between intrinsic and extrinsic measurements of bias [22].

Naturally occurring examples. NOEs are drawn from real-world corpora such as Wikipedia [14, 36, 70], Twitter [6, 49], or Reddit [4]. They preserve authentic linguistic properties, thereby capturing bias in specific domains. However, their availability is uneven across domains, their collection is costly, and they lack controllability, since semantics cannot be easily manipulated or systematically extended to represent different demographic groups. Thus, they pose a barrier to scalable bias evaluation.

Crowdsourced datasets. Crowdsourced datasets are constructed from bias-sensitive examples produced by human annotators. Representative resources include *StereoSet* and *CrowS-Pairs* [47, 48]. While these datasets offer flexibility and realism, they entail high financial costs.

In this work, we propose leveraging LLMs to adapt template-based resources to domain-specific contexts, combining the scalability and controllability of templates with the realism of naturally occurring text.

2.1.2 Names and Social Biases. Names are commonly used as proxies for demographic attributes in NLP bias evaluation. Prior work has adopted this approach to study biases related to *gender*, *race*, and their intersection across a range of settings, including word embeddings [10], sentence encoders [42], downstream classifiers [11, 31], commonsense reasoning [2, 27], and, more recently, LLM-based decision-making scenarios [1, 68].

Beyond these attributes, names have also been used to study *nationality*-related biases. Prior work has shown that names associated with less represented countries are more susceptible to biased model behavior [5, 6]. Names have also been used to evaluate intersectional biases involving gender and geographic attributes in generative models [32]. Complementary research has investigated the mechanisms underlying names and biases, showing that the underrepresentation of minority names in pretraining corpora is associated with poorer token coverage and weaker contextual representations [3, 72].

Associations between names and demographic groups have limitations and should therefore be used with caution. For example, Gautam et al. [21] note that these associations are often culturally specific, which can reduce construct validity and introduce cultural insensitivity. Likewise, Lockhart et al. [39] show that demographic inference from names can yield uneven error rates across groups. Still, names remain a practical tool for probing demographic disparities in model behavior. In our case, they are used as statistical proxies at the group level, not as reliable indicators of an individual's identity.

2.2 Style Transfer using LLMs

Recent work has shown that LLMs exhibit strong zero- and few-shot capabilities across a wide range of tasks [9, 34], including text style transfer [56, 66]. This task aims to modify stylistic attributes, such as sentiment or formality, while preserving meaning. LLMs have achieved strong results in style transfer according to both automatic metrics and human evaluation, without requiring parallel corpora or specialized resources [46].

Prior approaches include candidate generation with re-ranking [66], prompted generation of diverse styles [56], and prompt-based editing for finer stylistic control [40]. In this work, we draw on this line of research to rewrite bias-measurement templates so they better align with domain-specific language. We also adopt common style-transfer evaluation criteria, namely content preservation and domain adherence, to assess the adapted templates.

3 Method

This section presents the terminology used in our work, our method for adapting template-based bias datasets to specific domains using LLMs, as well as the evaluation framework used to assess its effectiveness.

3.1 Terminology

To establish the notation used in this work, we first introduce the main concepts. Following Czarnowska et al. [13], a **sensitive attribute** $\mathcal{S}$ (e.g., religion) refers to a social category that may be subject to bias. Each attribute is associated with a set of **protected groups** $\mathcal{G}$ (e.g., {*Christian, Muslim*}), and every group $G \in \mathcal{G}$ is characterized by a set of **identity terms** I_G (e.g., {*church, Bible*} for the group *Christian*).

Let M denote a model evaluated on a downstream task, and let $\mathcal{D}$ denote the application domain. We operationalize social bias as systematic differences in the predictions of M across protected groups when the non-demographic content of the input is held fixed. Accordingly, we define $B_{M,\mathcal{S},\mathcal{D}}$ as the bias exhibited by M in domain $\mathcal{D}$ with respect to attribute $\mathcal{S}$, considering its associated protected groups and identity terms.

Finally, we denote by $T = \{t^1, \ldots, t^N\}$ a collection of templates constructed to probe model bias with respect to $\mathcal{S}$, where each protected group $G \in \mathcal{G}$ is represented by identity terms from I_G.

3.2 LLM-Based Template Adaptation

Our objective is to adapt a set of templates T to more accurately estimate bias $B_{M,\mathcal{S},\mathcal{D}}$ in a target application domain $\mathcal{D}$. We assume access to a finite sample $D = \{d_1, \ldots, d_{|D|}\}$ drawn from real-world data in $\mathcal{D}$. For each template $t^i \in T$, we use a generator $g_{\mathcal{D}}$—an LLM prompted with in-domain examples—to rewrite t^i in the style of the target domain while preserving its variable slots and semantic meaning using examples from D. This produces a domain-adapted template $t^i_{\mathcal{D}} = g_{\mathcal{D}}(t^i)$.

The resulting collection of rewritten templates is defined as

$$T_{\mathcal{D}} = g_{\mathcal{D}}(T) = \{t^1_{\mathcal{D}}, \ldots, t^N_{\mathcal{D}}\},$$

yielding a dataset tailored to the linguistic characteristics of $\mathcal{D}$. Figure 2 illustrates this process.

We distinguish between the original template domain $\mathcal{D}_T$ and the adapted template domain $\mathcal{D}_{T_{\mathcal{D}}}$. Bias estimates can thus be obtained both from $B_{M,\mathcal{S},\mathcal{D}_T}$ and from $B_{M,\mathcal{S},\mathcal{D}_{T_{\mathcal{D}}}}$, enabling evaluation under conditions that more faithfully reflect real-world usage.

3.3 Evaluation of Bias Consistency

Since the goal of our domain-adapted datasets is to measure bias in models within a specific domain, we assess their effectiveness through *bias consistency*—comparing bias estimates from adapted

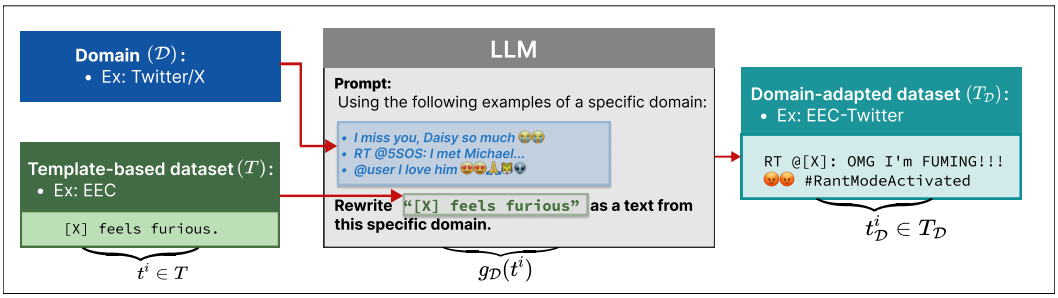


Fig. 2. Illustration of the domain-adaptation process. A set of template-based examples (e.g., EEC) is rewritten by an LLM using in-domain examples (e.g., Tweets) to generate domain-adapted templates. This procedure preserves the semantic slots of the original template while aligning its style with the target domain $\mathcal{D}$.

datasets with those obtained from NOEs in a model M trained in a downstream task. In our notation, this amounts to evaluating whether the set $T_{\mathcal{D}}$ provides a better approximation of bias $B_{M,\mathcal{S},\mathcal{D}}$ than the original set T.

Bias Estimation in the Application Domain. Estimating $B_{M,\mathcal{S},\mathcal{D}}$ for a given attribute $\mathcal{S}$ and domain $\mathcal{D}$ can be costly. To address this, we focus on settings where $\mathcal{S}$ can be represented through entities that act as proxies for social groups, following the approach used in Barriere and Cifuentes [6] and Naous et al. [49]. In this setting, constructing NOEs requires only a sample source from $\mathcal{D}$, from which we filter instances containing relevant entities via Named Entity Recognition (NER). These entities are then used to generate counterfactuals $N = \{n_1, \ldots, n_{|N|}\}$, that is, paired variants of the same sentence obtained by replacing the original entity with identity terms associated with different social groups while preserving the remaining context.

This strategy leverages naturally occurring data to approximate counterfactuals, making it possible to estimate bias automatically with relatively low overhead. However, its effectiveness is tied to the availability and reliability of named entities, and it restricts the notion of $\mathcal{S}$ to entity-based definitions, since $I_{\mathcal{G}}$ can only be instantiated through entity mentions. Even with these constraints, the method remains a practical way to approximate $B_{M,\mathcal{S},\mathcal{D}}$ and serves as a useful baseline for evaluating our approach.

Metrics. We adopt two metrics described in Czarnowska et al. [13]. First, the Vector Background Comparison Metric (VBCM), which encodes bias across protected groups as a vector, where the i-th component captures the disparity between the score of group $G_i \in \mathcal{G}$ and that of a designated reference group, referred to as the *background* β. Second, the Multi-group Comparison Metric (MCM), which captures the overall impact of a sensitive attribute $\mathcal{S}$ on model predictions across its associated protected groups.

Both metrics are based on two functions: ϕ, a scoring function that computes model outcomes over a subset of examples, and d, a comparison function that measures the discrepancy between sets of scores.

Bias Comparison. We then evaluate whether LLM-adapted templates $T_{\mathcal{D}}$ better capture real-world biases in domain $\mathcal{D}$ than the original templates T. To do so, we compare their VBCM vectors against those derived from naturally occurring examples N. For this purpose, we define a distance metric m_d that quantifies the difference between bias vectors across datasets. As shown in Figure 3, this allows us to assess whether adapted templates yield estimates that align more closely with NOEs than template-only baselines.

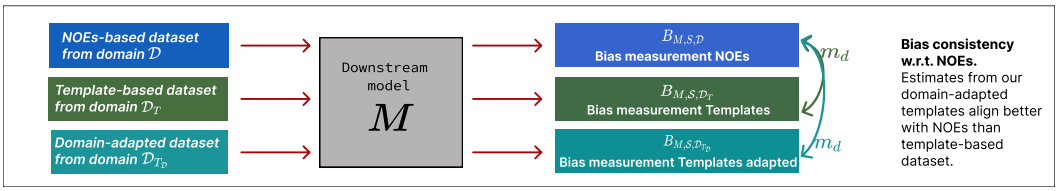


Fig. 3. Bias measurement criteria: we compare estimates from domain-adapted templates, template-based datasets, and NOEs. Domain-adapted templates yield bias estimates that align more closely with NOEs than template-only baselines.

4 Experimental Setup

This section presents the experimental setup. We describe the sensitive attributes, evaluation metrics, datasets, domains, and models used to assess the effectiveness of our proposed method.

4.1 Bias

4.1.1 Attributes. We focus on attributes that can be represented by multiple protected groups, as this enables meaningful comparisons of bias vectors across groups in the vector space. We also choose attributes that are well established in the NLP social bias literature. Following these criteria, we select **nationality**, which comprises 38 protected groups, and **gender-race/ethnicity**, which includes 8 protected groups.

Nationality bias. For this attribute, the protected groups $\mathcal{G}$ comprise 38 countries. Each group is represented by 50 popular names per country, following Barriere and Cifuentes [5, 6], who obtained these names from Wikidata Query Service and filtered them using Checklist.

Gender-race/ethnicity bias. We define eight groups based on two categories of gender {*female*, *male*} and four groups of race or ethnicity in the categories {*White*, *Black*, *Hispanic*, *Asian*}. Each group is represented by 50 names selected from the U.S. Social Security Administration (SSA) database[2]. We select names with more than 75% association with either female or male records in the SSA database and more than 75% association with one race/ethnicity group according to Rosenman et al. [58].

4.1.2 Bias measurement.

VBCM. Following Levy et al. [35], we define the *background value* as the average outcome across all protected groups. Building on this formulation, Zayed et al. [74] define the VBCM metric using two components: ϕ, the probability of a positive outcome, and d, demographic parity. Demographic parity is computed as $DP(x, y) = 1 - |x - y|$, where x denotes the positive outcome for a protected group and y corresponds to the outcome associated with the *background value*.

To assess the consistency of bias estimates across datasets, we compare the bias scores produced by the original and adapted template sets with those obtained from NOEs. We use two complementary metrics: mean absolute error (MAE), which captures the average difference between estimates, and Pearson correlation, which measures the extent to which bias patterns vary similarly across groups. Lower MAE indicates closer agreement with the NOE-based estimates, whereas higher Pearson correlation indicates stronger alignment in the overall bias patterns. Under this criterion, domain-adapted templates are preferred when they yield lower MAE and higher Pearson correlation than the original templates.

[2]https://www.ssa.gov/oact/babynames/background.html

MCM. MCM measures the variation in a model's predictions across demographic groups within a given attribute. To compute this metric, we use ϕ, the probability of a positive outcome, and d, the standard deviation, following Levy et al. [35] and Prabhakaran et al. [53]. MCM is always non-negative: a value of 0 indicates no variation in predictions across the groups of an attribute, whereas larger values indicate greater variation.

4.2 Template Datasets Studied

We apply our methodology to two template-based datasets originally created in English: the *Equity Evaluation Corpus* (EEC) and the *Identity Phrase Templates Test Set* (IPTTS). These datasets have been influential in the development of bias evaluation benchmarks and in research on bias mitigation for NLP models.

EEC, introduced by Kiritchenko and Mohammad [30], is designed to evaluate gender and racial biases in emotion regression models. It has also inspired recent multilingual bias studies [11]. In contrast, IPTTS, proposed by Dixon et al. [15], is a synthetic dataset for evaluating social bias in toxicity classification systems. Its design includes balanced toxic and non-toxic templates and has been widely used to assess and mitigate various types of biases [8, 15, 50, 74].

Since our sensitive attribute of interest is names, we use templates from EEC and IPTTS that can be perturbed using names. Example templates from these datasets are shown in Table 1.

Table 1. Example templates from the Equity Evaluation Corpus (left) and the Identity Phrase Templates Test Set (right).

EEC -Templates	**IPTTS** - Templates
I saw [PERSON] in the market.	*[PERSON] is an ugly nurse.*
[PERSON] feels happy.	*[PERSON] is a lovely fire fighter.*

4.3 Application Domains

We focus on three real-world domains $\mathcal{D}$: Twitter, Wikipedia Talk Pages and IMDb.

Twitter. Twitter is a microblogging platform where users post short messages (tweets) limited to 280 characters. Tweets are typically brief and informal, and may contain spelling or grammatical errors. The character limit encourages the use of abbreviations, acronyms, emojis, and internet slang [26], making Twitter a linguistically diverse and dynamic domain for studying bias in language models. This domain has previously been used to evaluate social bias with synthetic data [11, 43].

Wikipedia Talk Pages. The talk pages of Wikipedia are discussion forums where users coordinate edits, debate content, and interact around Wikipedia articles and user pages. These conversations are semi-structured and interactive, with an average length of 414 characters. The style is often informal and expressive, reflecting negotiation, disagreement, or collaboration among contributors. Such characteristics make this domain particularly challenging for domain adaptation and fairness evaluation. Prior work has leveraged this domain to study social bias and mitigation strategies using synthetic data [15, 50, 74].

IMDb. The Internet Movie Database is a widely used platform for publishing user-generated reviews of movies and television shows. Texts in this domain are generally longer than tweets, written in a narrative or opinionated style, and often reflect personal sentiments, preferences, and cultural perspectives.

Naturally Occurring Examples. We construct NOE-based datasets for each domain using EuroTweets [45] and SemEval-2018 for Twitter, the Wikipedia Talk Pages test set [73], and the IMDb Dataset of 50K Movie Reviews[3] [41] for IMDb.

Our framework requires sentences that support counterfactual substitutions across social groups. Since not all NOEs are suitable for this purpose, we retrieve in-domain sentences containing named entities and use them as counterfactual templates by replacing the entity with group-related identity terms. To identify suitable sentences, we follow the procedure of Barriere and Cifuentes [6], which uses SpaCy together with the Checklist library [57].

For Twitter, we obtain 1,000 examples in total: 451 from EuroTweets and 549 from SemEval-2018. For Wikipedia Talk Pages, we apply the same procedure to the full test set and obtain 1,101 examples. For IMDb, we use a sampled subset of the original 50K review corpus and retain 1,000 examples. We do not include more examples in this case because each sentence is expanded into multiple counterfactual variants, which increases the cost of large-scale evaluation.

4.4 Generated Templates

To obtain domain-adapted templates, we design a rewriting prompt that takes as input an original template containing an identity term and asks the LLM to preserve both the semantic meaning and the emotional tone of the original template while rewriting it in the linguistic style of the target domain, using $n = 15$ in-domain examples as contextual guidance. The complete prompt is shown in Figure 4.

For the target domains, we draw in-domain examples from the EuroTweets corpus, the Wikipedia Talk Pages training set, and the IMDb 50K Movie Reviews dataset. A distinctive feature of our approach is that the examples used for prompting can be arbitrary sentences sampled from the target domain, unlike NOEs, which must explicitly contain named entities. To make comparisons across datasets fairer, we exclude NOEs from the prompting pool, so that the adaptation process does not rely on the same examples later used as reference during evaluation.

We generate domain-adapted templates using three instruction-tuned large language models: `LLaMA3.3-70B-Instruct` (`LLaMA3.3-70B`), `LLaMA3.1-8B-Instruct` (`LLaMA3.1-8B`) [16], and `Mixtral-8x7B-Instruct-v0.1` (`Mixtral8x7B`) [28]. Additional details of the generation process are provided in Appendix A.

4.5 Downstream Models

To evaluate the different social bias datasets, we use models trained for the corresponding downstream tasks.

For EEC, we follow the original benchmark setting and use the SemEval-2018 Task 1 dataset [44] to train the downstream sentiment models. Specifically, we fine-tune three encoder-based models, `BERT`, `DistilBERT`, and `cardiffnlp/twitter-roberta-base`, as well as one LLM-based classifier, `LLaMA3.1-8B`. We also include a popular off-the-shelf classifier `Cardiffnlp-sentiment`[4].

For IPTTS, we similarly follow the original benchmark setup and use the Wikipedia Talk corpus [15] to train the downstream toxicity models. In this case, we fine-tune `BERT`, `DistilBERT`, and `LLaMA3.1-8B`. In addition, we evaluate two off-the-shelf classifiers: `Cardiffnlp-hate`[5] and `Cardiffnlp-offensive`[6].

[3] https://www.kaggle.com/datasets/lakshmi25npathi/IMDB-Dataset-of-50K-movie-reviews
[4] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-hate
[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

```
Prompt

I want you to rewrite a text in the style of a specific domain.
The following [n] texts are examples of this specific domain:
 1. [example 1]
…
 n. [example n]
Can you rewrite the following sentence [template(identity-term)]
as domain-specific text?
This means that the new text must retain the general emotion and
maintain the semantic meaning of the original sentence, but use
the style of the specific domain.
You should also include [identity-term] explicitly only once in
the new text.
Output only the rewritten text, without any introduction or
explanation.
```

Fig. 4. Prompt used for LLM-based template adaptation.

4.6 Hardware and Implementation Details

All experiments were implemented in PyTorch using Hugging Face Transformers and run on a cluster with NVIDIA RTX A6000 GPUs (48 GB), using up to two GPUs per run when needed. For large generative models, we used quantization and automatic device placement to fit them within the available hardware.

For template adaptation, LLaMA3.1-8B was loaded in 8-bit precision, while LLaMA3.3-70B and Mixtral8x7B were loaded with 4-bit quantization. Generation used a temperature of 0.01, and the maximum length was set dynamically as the input length plus the mean tokenized length of in-domain examples.

For downstream discriminative models, we used Hugging Face classifier backbones. Masked language models were fine-tuned with AdamW, a learning rate of 5×10^{-5}, batch size 16, and 3 epochs, using a linear scheduler without warm-up. For the LLaMA3.1-8B regression model, we fine-tuned a 4-bit quantized sequence-classification variant with LoRA ($r = 8$, $\alpha = 16$, dropout 0.05) on the q_proj and v_proj modules. Inputs were truncated to 128 tokens, and training used an effective batch size of 16, a learning rate of 1×10^{-4}, and 5 epochs.

5 Results

In this section, we present the results of our experiments. We begin by visualizing VBCM bias vectors to examine how closely the LLM-adapted templates approximate the values obtained from NOEs. We then report distance- and correlation-based metrics to quantitatively assess the effectiveness of the proposed method, followed by an aggregated comparison across domains and downstream models. Next, we present the analysis based on the MCM metric. Finally, we evaluate the sensitivity of the adaptation framework to prompt design by comparing the default setting with a few-shot variant and by varying the number of in-domain examples provided during adaptation.

5.1 VBCM Visualization

Figure 5 presents the VBCM values for nationality groups in the Twitter domain using IPTTS, comparing the original template-based dataset, its LLM-adapted variants, and NOEs across three

downstream models: the fine-tuned classifiers `DistilBERT (ft)` and `LLaMA3.1-8B (ft)`, and the off-the-shelf model `Cardiffnlp-offensive`.

Across all three models, the original template-based dataset (green) departs more substantially from the NOE-based estimates (red), indicating that it produces larger bias estimates than the NOE-based reference, leading to weaker correlations with real-world data in the three different models. This pattern is consistent with the findings of Kaneko et al. [29], who observed that artificial templates often amplify gender biases compared to natural sentences.

Importantly, this gap is reduced when using IPTTS domain-adapted datasets generated by LLMs (blue), which produce estimates that more closely approximate NOE values. This demonstrates the effectiveness of LLM-based adaptation in aligning synthetic bias measurements with those obtained from authentic data. Furthermore, the magnitude of these differences varies across models, suggesting that each downstream model's sensitivity to domain shift plays a role in how well synthetic and real-world bias estimates align.

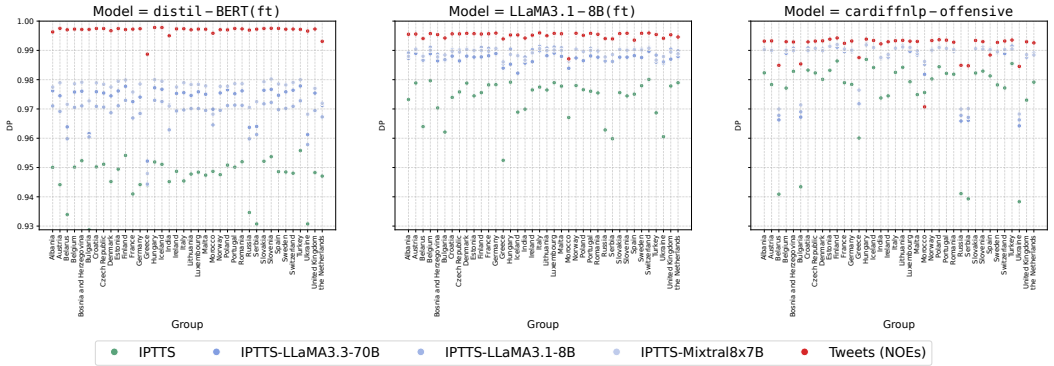


Fig. 5. Demographic Parity (DP) scores across nationality groups for three classification models (DistilBERT, LLaMA3.1-8B and Cardiffnlp-Offensive). Results are shown for the IPTTS dataset adapted with different LLMs (LLaMA3.3-70B, LLaMA3.1-8B, Mixtral8x7B) and compared against naturally occurring examples (NOEs) from Tweets.

5.2 VBCM Distance and Correlation

In this subsection, we quantitatively evaluate the alignment between bias estimates derived from template-based datasets and those obtained from NOEs. Specifically, we compare VBCM vectors using mean absolute error (MAE) and Pearson correlation across multiple downstream models. We then aggregate the results by LLM, domain, and dataset to identify the configurations that yield the most consistent performance in bias measurement.

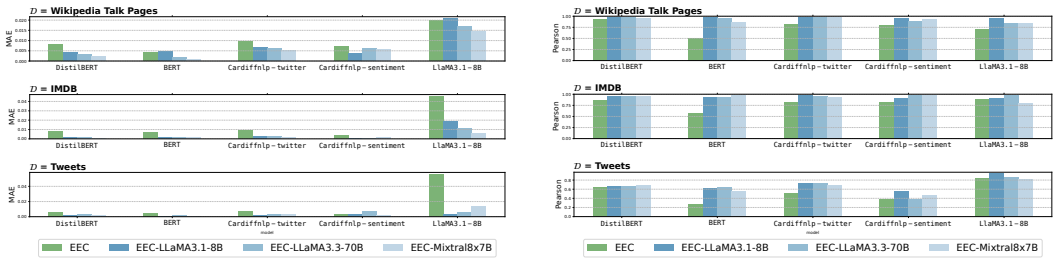
5.2.1 EEC. In Figure 6, we report a comparison with the VBCM metrics obtained with the EEC dataset and its domain-adapted variants, compared against NOEs. Results are presented using MAE (Figure 6a) and Pearson correlation (Figure 6b) for the attribute *nationality* across the different downstream models.

Across models, we observe differences in MAE and Pearson correlation between the template-based datasets and NOEs, although the magnitude of these differences varies across models. For MAE, the fine-tuned model `LLaMA3.1-8B` shows the largest differences across domains, with values of 0.050, 0.020, and 0.046 in Twitter, Wikipedia Talk Pages and IMDb, respectively. In Pearson correlation, the `BERT` model exhibits the widest discrepancies between EEC and NOEs,

with differences of 0.67, 0.499, and 0.566 across the three domains. These substantial gaps highlight the risk of relying solely on synthetic template-based datasets for bias measurement, as they may diverge considerably from real-world examples.

At the same time, a clear trend emerges: domain-adapted templates generated with LLMs tend to yield lower MAE and higher correlation with NOEs than the original EEC, indicating stronger alignment with real-world data.

5.2.2 IPTTS. Figure 7 shows the corresponding analysis for the IPTTS dataset. Results are broadly consistent with those of EEC, although the extent of improvement differs across models. These variations may reflect different levels of robustness to domain shift, which influence the degree of discrepancy between template-based datasets and NOEs. In the case of IPTTS, the largest differences in both MAE and Pearson correlation are observed with the `DistilBERT` model, reinforcing that the impact of domain adaptation is highly model-dependent.



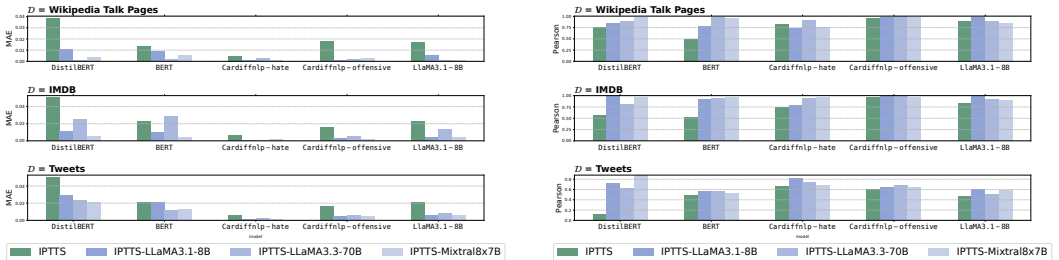
(a) MAE EEC (lower is better)

(b) Pearson EEC (higher is better)

Fig. 6. VBCM scores for Nationality on EEC across downstream models.

5.2.3 Overall Results. Table 2 summarizes the extent to which LLM-adapted templates yield more consistent bias measurements compared to their original counterparts. The table reports average differences across models and their aggregated values across domains.

To give a concrete example: a more positive difference in correlation (Δr) indicates that the LLM-adapted template achieves higher alignment with the NOEs than the original template. Conversely, a more negative difference in error (ΔMAE) means that the adapted template exhibits smaller discrepancies with respect to NOEs than the original version. For instance, a value of -0.011 for Mixtral8×7B on Twitter within the EEC dataset indicates that the distance between NOEs and EEC–Mixtral8×7B outputs is 0.011 lower than the distance between NOEs and the original EEC



(a) MAE IPTTS (lower is better)

(b) Pearson IPTTS (higher is better)

Fig. 7. VBCM scores for Nationality on IPTTS across downstream models.

templates. In the same vein, an overall ΔPearson value of 0.135 means that the Pearson correlation between the adapted-template and NOE bias estimates is 0.135 higher than that obtained with the original templates.

Results are reported for two sensitive attributes, *nationality* and *gender-race/ethnicity*, using both the EEC and IPTTS datasets.

Building on these results, the LLM models show overall improvements across domains and datasets when averaged over all evaluated models. Among them, Mixtral8×7B emerges as the most consistent performer across both attributes and datasets. On average, IPTTS achieves larger gains compared to EEC. Overall, IMDb tends to show the strongest improvements, whereas Wikipedia Talk Pages generally shows smaller gains.

Table 2. Average differences across models in how much domain-adapted templates improve bias measurement relative to original templates, evaluated against NOEs. Results are reported for two attributes (**Nationality** and **Gender-Race/Ethnicity**), three LLMs, and two metrics (Pearson r, MAE) across domains. Higher difference ΔPearson r and lower difference in ΔMAE indicate stronger agreement with NOEs for domain-adapted templates. **Mean** indicates the average across domains.

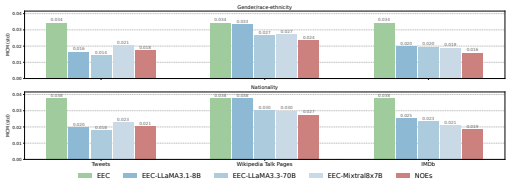
			EEC				IPTTS			
Attribute	**Metric**	**Model**	**Twitter**	**WTP**	**IMDb**	**Mean**	**Twitter**	**WTP**	**IMDb**	**Mean**
Nationality	ΔMAE	LLaMA3.1-8B	-0.012	-0.002	-0.010	-0.008	-0.011	-0.013	-0.018	-0.014
		LLaMA3.3-70B	-0.010	-0.003	-0.011	-0.008	-0.013	-0.011	-0.011	-0.012
		Mixtral8×7B	-0.011	-0.004	-0.013	**-0.009**	-0.014	-0.016	-0.021	**-0.017**
	ΔPearson r	LLaMA3.1-8B	0.131	0.223	0.147	**0.167**	0.171	0.113	0.206	0.163
		LLaMA3.3-70B	0.108	0.178	0.166	0.151	0.191	0.147	0.195	0.178
		Mixtral8×7B	0.135	0.164	0.178	0.159	0.227	0.168	0.227	**0.201**
Gender-Race/Ethnicity	ΔMAE	LLaMA3.1-8B	-0.012	-0.002	-0.011	-0.008	-0.011	-0.013	-0.016	-0.013
		LLaMA3.3-70B	-0.012	-0.003	-0.012	-0.009	-0.014	-0.014	-0.012	-0.013
		Mixtral8×7B	-0.011	-0.005	-0.012	**-0.009**	-0.013	-0.015	-0.021	**-0.017**
	ΔPearson r	LLaMA3.1-8B	0.186	0.179	0.142	0.169	0.198	0.110	0.396	0.233
		LLaMA3.3-70B	0.165	0.010	0.148	0.137	0.231	0.180	0.181	0.197
		Mixtral8×7B	0.244	0.183	0.132	**0.186**	0.268	0.301	0.377	**0.319**

5.3 MCM

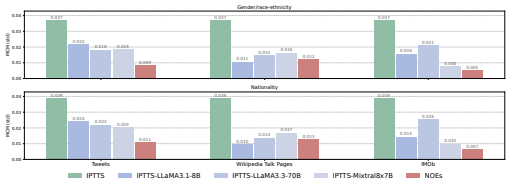
Figure 8 reports MCM as a complementary view of the variation in model predictions across demographic groups. Lower values indicate less variation across groups. Across both EEC and IPTTS, the original template-based datasets consistently yield the highest dispersion, whereas the LLM-adapted variants substantially reduce it and move closer to the NOE-based reference. This pattern is observed for both nationality and gender-race/ethnicity, indicating that the adapted templates better approximate the level of between-group variation observed in NOEs. The gains are generally larger for IPTTS than for EEC, with IMDb showing the strongest improvements and Wikipedia Talk Pages the smallest. Among the generator models, Mixtral8x7B is often the closest to the NOE values, indicating the most consistent overall behavior.

5.4 Prompt Sensitivity

We assess the sensitivity of our framework to prompt design through two complementary analyses. First, we compare the original zero-shot adaptation prompt against a few-shot variant while keeping the number of in-domain examples fixed at $n = 15$. Here, we use the term *zero-shot* to denote that the prompt does not include input-output demonstrations of the rewriting task, even though it provides in-domain examples as contextual guidance. Second, we examine the effect of varying the number of in-domain examples provided to the generator. Together, these analyses allow us



(a) MCM (std) EEC



(b) MCM (std) IPTTS

Fig. 8. MCM scores across datasets, target domains, and bias attributes. Lower values indicate lower variation in model predictions across demographic groups.

to determine whether the improvements observed in the main experiments depend on a specific prompt configuration.

5.4.1 Few-shot demonstrations. In this setting, we augment the original adaptation prompt with three rewritten examples that illustrate the desired transformation, while keeping the number of in-domain examples fixed at $n = 15$. We keep the generator models unchanged and regenerate the adapted templates, evaluating their agreement with NOEs for EEC under both bias settings, *Nationality* and *Gender-Race/Ethnicity* in two downstream models BERT and *DistilBERT*.

Table 3 compares the original zero-shot prompt and the few-shot variant across domains, bias settings, and generator models. Overall, adding few-shot demonstrations does not lead to consistent improvements. In most settings, the original prompt yields equal or better agreement with NOEs, particularly in terms of Pearson correlation. The strongest advantages of the original prompt appear in Wikipedia Talk Pages and IMDb, whereas the differences are smaller and more mixed for Twitter. The MAE results follow a similar pattern, although the magnitude of the differences is generally modest.

Overall, these results suggest that additional prompting guidance is not necessarily beneficial for domain-adaptive bias evaluation. In our setting, the original prompt provides a better balance between domain adaptation and semantic fidelity.

Table 3. Differences in agreement with NOEs between the original zero-shot prompt and the few-shot setting, with the number of in-domain examples fixed at $n = 15$. Positive values for ΔPearson r and negative values for ΔMAE favor the original prompt.

		Nationality			Gender-Race/Ethnicity		
Metric	**Model**	**Twitter**	**WTP**	**IMDb**	**Twitter**	**WTP**	**IMDb**
ΔMAE	LLaMA3.1-8B	-0.0002	-0.0045	0.0004	-0.0008	-0.0037	-0.0003
	LLaMA3.3-70B	0.0012	-0.0068	0.0005	0.0007	-0.0053	-0.0002
	Mixtral8x7B	-0.0004	-0.0039	-0.0005	-0.0014	-0.0036	-0.0006
ΔPearson r	LLaMA3.1-8B	0.0268	0.0447	0.0073	0.0312	0.0543	0.1166
	LLaMA3.3-70B	0.0352	0.0326	-0.0260	-0.0016	0.0753	0.1177
	Mixtral8x7B	-0.0046	-0.0284	0.0328	-0.0006	0.0372	0.0706

5.4.2 Effect of the number of in-domain examples. We next analyze the effect of varying the number of in-domain examples included in the adaptation prompt. Figure 9 reports agreement with NOEs as n increases across domains and generator models. Overall, performance remains relatively stable, with no uniform monotonic relationship between larger values of n and better agreement with NOEs.

On Twitter, both MAE and Pearson correlation vary only moderately across values of n, suggesting that performance is relatively insensitive to the prompting budget in this domain. On IMDb,

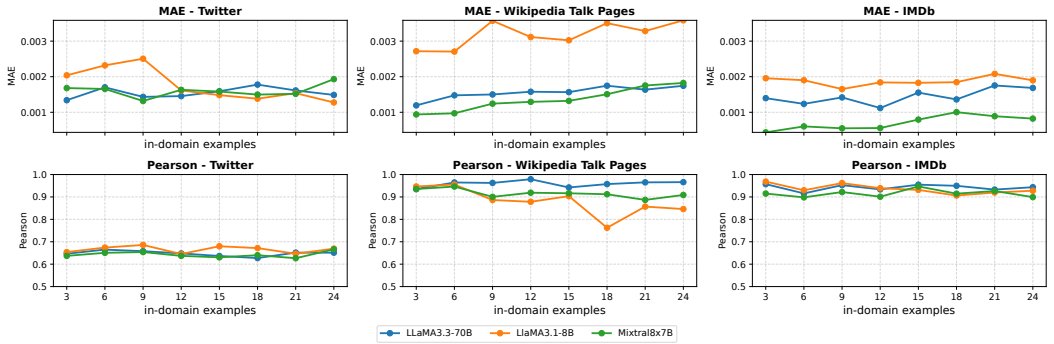


Fig. 9. Agreement with NOEs as a function of the number of in-domain examples in the adaptation prompt. Results are shown for MAE and Pearson r across target domains and generator LLMs. Lower MAE and higher Pearson r indicate better agreement.

all generator models maintain strong and stable performance across the full range, indicating diminishing returns once a small number of in-domain examples is provided. Wikipedia Talk Pages exhibits the largest variability, especially for `LLaMA3.1-8B`, which suggests that this domain is more sensitive to prompt composition. Even so, the overall ranking of generator models remains broadly similar across values of n.

Taken together, these results indicate that our framework is reasonably robust to the number of in-domain examples used in the prompt. While increasing n can improve agreement with NOEs in some settings, the gains are not systematic. The main configuration with $n = 15$ therefore provides a reasonable trade-off between performance and prompt length.

6 In-depth Domain Translation Analysis: Content Preservation and Style Transfer

To comprehensively evaluate our method for generating domain-adapted bias datasets, we assessed the LLM-generated templates using two criteria drawn from the style transfer literature: content preservation (Section 6.1.1) and domain adherence (Section 6.1.2) (see Figure 10). The following sections detail the experiments conducted for each evaluation criterion.

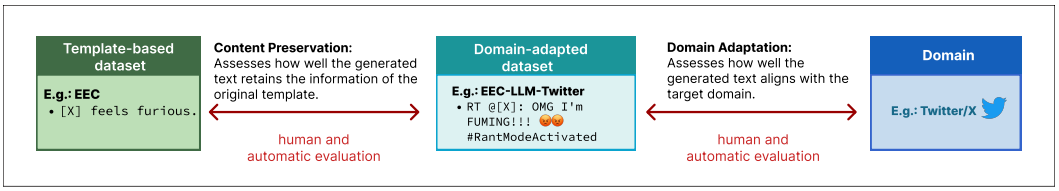


Fig. 10. Style transfer criteria: adaptation quality is assessed through content preservation (semantic similarity to the original templates) and domain adherence (linguistic fit to the target domain).

6.1 Evaluation Methodologies

To assess the quality of the domain-adapted templates, we designed an evaluation protocol inspired by the style transfer literature. Our aim is both to verify that the adapted templates preserve the intended bias structure and to ensure that they resemble realistic, domain-appropriate sentences. We therefore focus on two complementary criteria: *content preservation* and *domain transfer strength*, each evaluated through a combination of automatic metrics and human judgments.

6.1.1 Content Preservation. The original templates are carefully curated sentences that control how demographic groups are represented in specific linguistic contexts, for example, when sentences involve emotions. When LLMs rewrite these templates into domain-adapted versions, it is essential that the intended semantic meaning is preserved. To measure this, we conducted both automatic and human evaluations.

Automatic Evaluation. Inspired by Fu et al. [17], we measured semantic similarity between the original and adapted sentences using embedding-based representations. Specifically, we applied the Sentence Transformers library[7] to compute cosine similarity between sentence pairs. Higher similarity scores indicate stronger preservation of the original content.

Human Evaluation. To complement the automatic metrics, three independent annotators (with advanced academic training and no conflicts of interest) evaluated 225 adapted sentences from the EEC dataset. Following Li et al. [37], annotators rated the degree of meaning preservation on a Likert scale from 1 (completely dissimilar) to 5 (completely similar).

6.1.2 Domain Transfer Strength. Even if adapted templates produce bias estimates consistent with NOEs, it is also necessary to confirm that LLMs reproduce the stylistic and linguistic features of the target domain. Evaluating domain transfer strength ensures that the resulting datasets are faithful to real-world usage. This dimension was also assessed with both automatic and human evaluations.

Automatic Evaluation. We adopted the approach of Prabhumoye et al. [54], training a domain classifier on naturally occurring examples from each of the three domains. The proportion of adapted sentences correctly classified as belonging to the target domain was used as the transfer strength metric. Higher accuracy reflects stronger alignment with domain-specific linguistic features.

Human Evaluation. Inspired by Gong et al. [23], annotators were presented with sentence pairs: one from the target domain and another either generated by an LLM or sampled from a different domain. They were asked to judge which sentence was more likely to belong to the target domain. Transfer strength was calculated as the percentage of adapted sentences chosen as representative of the target style. For comparison, 25 naturally occurring sentences per domain were also included.

6.2 Evaluation Results

6.2.1 Automatic evaluation. Table 4 reports the automatic evaluation results for content preservation and domain transfer. Overall, the adapted sentences exhibit moderate similarity to the original templates. Although absolute similarity values are not particularly high, this is expected since the similarity model compares sentences across different domains, which reduces its effectiveness. This effect is most evident in the Twitter domain, where larger linguistic variations result in lower similarity scores. Across models, performance remains relatively consistent, with IPTTS achieving higher content preservation than EEC, likely due to its simpler sentence structures. This trend is consistent with the bias consistency analysis against NOEs.

For domain transfer, both IMDb and Twitter consistently achieve high scores across models. In contrast, Wikipedia Talk Pages show lower performance, particularly with Mixtral8×7B, while the LLaMA models reach moderate-to-high levels. This indicates that Wikipedia Talk style is inherently more difficult to capture, likely because it is less distinctive and more general than other domains. Moreover, as shown in Table 5, LLaMA models tend to overemphasize domain-specific markers in their generations—such as frequent use of emojis, hashtags, and user mentions in the Twitter domain. This exaggerated stylistic adaptation may partly account for their higher automatic transfer scores, even if it does not always align with natural in-domain usage.

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 4. Automatic evaluation results for Content Preservation and Domain Transfer Strength across domains (IMDb, Twitter, Wikipedia Talk Pages) and two benchmarks (EEC, IPTTS). Results are reported for three LLMs. Bold values indicate the best performance within each domain.

		Content Preservation		**Domain Transfer Strength**	
Domain	LLM	EEC	IPTTS	EEC	IPTTS
Twitter	LLaMA3.1-8B	0.528	0.682	0.958	0.857
Twitter	LLaMA3.3-70B	0.553	0.643	1.000	0.938
Twitter	Mixtral8x7B	0.583	0.668	0.896	0.767
WTP	LLaMA3.1-8B	0.648	0.635	0.701	0.821
WTP	LLaMA3.3-70B	0.654	0.696	0.542	0.771
WTP	Mixtral8x7B	0.636	0.669	0.285	0.413
IMDb	LLaMA3.1-8B	0.662	0.690	0.764	0.799
IMDb	LLaMA3.3-70B	0.692	0.766	0.833	0.816
IMDb	Mixtral8x7B	0.635	0.742	0.757	0.947

Table 5. Average number of special characters and emojis in domain-adapted templates generated by each LLM.

LLM	**n @**	**n #**	**n Emojis**
LLaMA3.1-8B	0.778	1.375	0.938
LLaMA3.3-70B	0.938	0.854	2.285
Mixtral8×7B	0.021	0.757	0.625

6.2.2 Human evaluation. Table 7 reports the human evaluation results for EEC. Across all models and domains, annotators assigned high content-preservation scores, indicating that the adapted sentences generally retained the meaning of the original templates. In addition, human judgments on domain transfer strength broadly mirror the automatic evaluation results: adaptations to Tweets and IMDb were identified with high accuracy, whereas WTP was the least consistently recognized domain (see Appendix B for inter-annotator agreement).

A likely reason why WTP adaptations are less readily recognized is that, among the target domains, WTP is the most lexically and discursively distant from the source templates. EEC consists largely of explicit emotion- and sentiment-oriented statements, and although this type of affective language is also more often present in IMDb and Twitter, it is less characteristic of WTP, which is more often shaped by article-related discussion, disagreement, coordination, and meta-commentary. As a result, WTP rewrites may preserve the original meaning while still retaining traces of the source affective framing, making the target-domain style less immediately identifiable to annotators.

This explanation is supported by both the lexical overlap analysis and the annotation results. Table 6 shows that WTP has the lowest normalized lexical overlap with EEC among the target domains, indicating that it is the most lexically distant from the source templates.

The annotation results follow the same pattern. Annotators were more likely to confuse WTP adaptations with IMDb than with Twitter, with error rates of 41% and 10%, respectively. This suggests that WTP adaptations are easier to distinguish when contrasted with Twitter, likely because Twitter contains more salient stylistic cues. In contrast, the distinction becomes less clear

Table 6. Normalized lexical overlap between the EEC vocabulary and naturally occurring texts from each target domain, excluding stopwords. Lower values indicate a greater lexical distance from the source templates.

Domain	Normalized lexical overlap with EEC
Twitter	0.0079
WTP	0.0038
IMDb	0.0099

when WTP adaptations are compared with IMDb sentences, since both can express emotions or personal reactions, even if neither fully matches the discourse structure typical of WTP.

A qualitative example illustrates this difficulty. The WTP adaptation of *"The situation makes Ryan feel discouraged."* becomes *"' Ryan contemplates the current circumstances, which leave him feeling disheartened."* Although this rewrite includes some cues compatible with WTP, none of the annotators identified it as belonging to that domain.

Table 7. Human evaluation results for Content Preservation (1–5 Likert scale) and Domain Transfer Strength (proportion of correct classifications) across domains (Twitter, Wikipedia Talk, IMDb) for three LLMs. Mean scores are reported, and NOEs are included as reference points.

	Content Preservation			**Domain Transfer Strength**		
Model	Twitter	Wikipedia Talk	IMDb	Twitter	Wikipedia Talk	IMDb
LLaMA3.1-8B	4.06 ± 1.186	4.01 ± 1.299	3.93 ± 1.027	0.98 ± 0.067	0.67 ± 0.373	0.97 ± 0.092
LLaMA3.3-70B	4.30 ± 0.838	3.74 ± 1.284	3.91 ± 1.324	0.97 ± 0.092	0.81 ± 0.306	0.96 ± 0.147
Mixtral8x7B	4.06 ± 0.491	4.25 ± 0.491	4.48 ± 0.491	0.93 ± 0.236	0.55 ± 0.371	0.88 ± 0.270
Mean	4.15 ± 0.930	4.00 ± 1.172	4.11 ± 1.029	0.96 ± 0.151	0.68 ± 0.363	0.94 ± 0.187
NOEs	–	–	–	0.97 ± 0.133	0.98 ± 0.067	1.00 ± 0.000

7 VBCM Geometric Analysis

This section presents a final analysis based on geometry in the VBCM vector space. Specifically, we examine the position of the adapted template vector $VBCM_{T_{\mathcal{D}}}$ relative to both the original template vector $VBCM_T$ and the naturally occurring examples $VBCM_{NOEs}$. The goal is to determine whether $VBCM_{T_{\mathcal{D}}}$ acts as an interpolation between synthetic templates and real-world examples from domain $\mathcal{D}$.

To this end, we employ two complementary measures. First, we compute the cosine similarity between $VBCM_{T_{\mathcal{D}}}$ and $VBCM_{NOEs}$, which indicates the degree of collinearity between the direction of adaptation (from templates to NOEs) and the adapted template vector. Second, we project $VBCM_{T_{\mathcal{D}}}$ onto the line connecting $VBCM_T$ and $VBCM_{NOEs}$, and calculate the relative distance ratio. A value close to 0 means that the adapted vector remains near the original templates, while a value close to 1 indicates stronger directional alignment to the NOEs vector. An overview of this procedure is shown in Figure 11, and the results are summarized in Table 8.

The results show consistent trends across both attributes and datasets. For nationality, Mixtral8×7B achieves the strongest performance, with cosine similarities above 0.93 in both EEC and IPTTS, and projection distances closer to NOEs (0.37 and 0.649). LLaMA3.3-70B and LLaMA3.1-8B also yield competitive cosine similarities (0.863–0.923), though their projections remain closer to the template side.

Table 8. Comparison of cosine similarity (Cos. Sim.) and projection distance (Proj. Dist.) for EEC and IPTTS across attributes, averaged across domains. Values closer to 1 indicate that the adapted bias vector is closer to the NOE-based bias vector.

Attribute	Model	EEC		IPTTS	
		Cos. Sim.	Proj. Dist.	Cos. Sim.	Proj. Dist.
Nationality	LLaMA3.1-8B	0.863	0.371	0.923	0.633
	LLaMA3.3-70B	0.904	**0.371**	0.893	0.512
	Mixtral8x7B	**0.930**	0.370	**0.993**	**0.649**
Gender-Race/Ethnicity	LLaMA3.1-8B	0.951	**0.083**	0.991	0.122
	LLaMA3.3-70B	0.931	0.082	0.986	0.112
	Mixtral8x7B	**0.962**	0.075	**0.998**	**0.133**

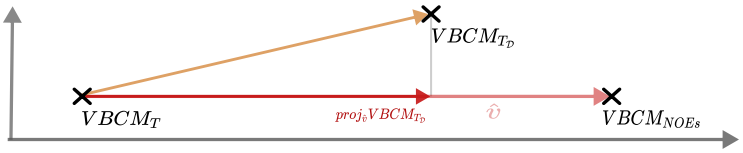


Fig. 11. Geometric analysis in the 2D VBCM space, evaluating the position of $VBCM_{T_\mathcal{D}}$ relative to $VBCM_T$ and $VBCM_{NOEs}$ through cosine similarity and vector projection.

For gender-race/ethnicity, all models exhibit very high cosine similarity, indicating that the adapted vectors move in a direction closely aligned with the template–NOE direction. However, the relatively low projection ratios (0.075–0.133) show that the adapted vectors remain close to the original templates. In contrast, the nationality setting exhibits larger projection ratios, particularly for IPTTS, indicating greater progress toward the NOE position.

Overall, these findings show that LLM-based adaptation generally moves the bias vectors in a direction aligned with the NOE-based estimates. However, the magnitude of this movement varies across attributes: adaptation produces a larger positional shift toward NOEs for nationality, whereas the gender-race/ethnicity vectors remain closer to the original templates despite their strong directional alignment.

8 Conclusion

Domain plays a central role in social bias measurement. We find that bias estimates systematically depend on whether evaluation datasets are domain-sensitive or domain-agnostic. This result is consistent across model types and adaptation settings, including off-the-shelf, fine-tuned generative models and masked language models.

To address this problem, we adapt template-based bias benchmarks to target domains using LLMs. We evaluate this approach on two benchmark datasets, EEC and IPTTS, along three dimensions: bias consistency, content preservation, and domain transfer. Domain-adapted templates align more closely with NOEs at both the attribute and group levels. At the same time, human and automatic evaluations show that the rewrites largely preserve the semantics of the original templates. Domain transfer is strongest for Tweets and IMDb, while WTP remains the most difficult setting, likely because it is lexically and topically farther from the source templates.

These results highlight domain as a key factor in bias evaluation. Ignoring domain context can distort fairness estimates and lead to misleading conclusions about mitigation methods. In practical

evaluation workflows, developers should therefore complement general-purpose bias benchmarks with domain-adapted evaluations that better reflect the target deployment setting. More broadly, our framework shows that open-source LLMs can be used to build domain-adapted bias resources that retain the controllability of templates while better reflecting naturally occurring in-domain language.

Future work should focus on improving adaptation methods for discourse-heavy and structurally distant domains, analyzing alternative prompting strategies, and developing new ways to compare results against larger NOE sets. It would also be valuable to extend the framework to low-resource settings, languages beyond English, and additional NLP tasks.

9 Limitations

Although our framework is a step toward domain-sensitive bias evaluation, it has several limitations.

First, our validation against NOEs relies on a pipeline that retrieves in-domain sentences containing named entities and filters them for counterfactual suitability. While this design allows for comparison with examples that more accurately reflect the target domain, it also introduces uncertainty regarding the reliability and scope of the NER-based extraction and filtering process. Consequently, the selected NOEs are restricted to a subset of instances that satisfy these constraints and may not accurately represent the domain. Future work should explore alternative forms of NOEs, more robust retrieval strategies, and manual validation to improve coverage and strengthen the robustness of this reference.

Second, our experiments are limited to English, operationalize gender using two categories, consider four race/ethnicity groups within a U.S.-specific context, and represent social groups through personal names. These choices provide only a partial view of social identity and may not capture more complex or context-dependent forms of bias. However, the core contribution of our methodology is to propose a framework for constructing domain-sensitive bias evaluation datasets. In this sense, names are used as an instrumental mechanism for instantiating social groups, rather than as a comprehensive representation of identity. We leave to future work the evaluation of adapted templates using other types of identity terms, such as nouns, pronouns, or explicit group descriptors.

Third, although LLM-based generation is effective for adapting templates across domains, it remains susceptible to errors, such as hallucinations and exaggerated stylistic artifacts. While our content-preservation evaluation helps identify some of these issues, such effects may persist in the adapted templates. Furthermore, although our approach relies on open-source LLMs, it still requires substantial computational resources for inference. This may limit accessibility in low-resource settings and entail environmental costs.

Finally, as with any bias evaluation framework, the resulting scores should not be interpreted as definitive measures of fairness.

Acknowledgments

This work was funded by U-INICIA 2024 from the Vicerrectoría de Investigación y Desarrollo (VID), grant UI-011/24, “Estudios de sesgos sociales en modelos de lenguajes largos” (TQ, VB); the ANID Chile National Doctoral Scholarship 21220586 (TQ); the National Center for Artificial Intelligence (CENIA), FB210017, Basal ANID (TQ, VB, FB); the ANID Chile Millennium Science Initiative Program, IMFD ICN17_002 (TQ, FB); and FONDEF IDeA I+D 2025, grant ID25I10330 (FB). We thank the anonymous reviewers for their insightful comments and suggestions, which helped improve this work.

References

[1] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? *arXiv preprint arXiv:2406.10486* (2024).

[2] Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2022. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. *arXiv preprint arXiv:2210.07269* (2022).

[3] Haozhe An and Rachel Rudinger. 2023. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. *arXiv preprint arXiv:2305.16577* (2023).

[4] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521* (2021).

[5] Valentin Barriere and Sebastian Cifuentes. 2024. Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method with Least Confounding Variables. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1511–1518.

[6] Valentin Barriere and Sebastian Cifuentes. 2024. A Study of Nationality Bias in Names and Perplexity using Off-the-Shelf Affect-related Tweet Classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 569–579. doi:10.18653/v1/2024.emnlp-main.34

[7] Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. *arXiv preprint arXiv:2010.14534* (2020).

[8] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*. 491–500.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[11] António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the Multilingual Margins: Intersectional Biases of Sentiment Analysis Systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar (Eds.). Association for Computational Linguistics, Dublin, Ireland, 90–106. doi:10.18653/v1/2022.ltedi-1.11

[12] Marta R Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. *arXiv preprint arXiv:2305.13198* (2023).

[13] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics* 9 (2021), 1249–1267.

[14] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.

[15] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[17] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[18] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

[19] Lance Calvin Lim Gamboa, Yue Feng, and Mark Lee. 2025. Social Bias in Multilingual Language Models: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 27845–27868.

[20] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences* 11, 7 (2021), 3184.

[21] Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. *arXiv preprint arXiv:2405.17159* (2024).

[22] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859* (2020).
[23] Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. *arXiv preprint arXiv:1903.10671* (2019).
[24] Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J Passonneau. 2023. Sociodemographic bias in language models: A survey and forward path. *arXiv preprint arXiv:2306.08158* (2023).
[25] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100* (2020).
[26] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016).
[27] Sullam Jeoung, Jana Diesner, and Halil Kilicoglu. 2023. Examining the Causal Impact of First Names on Language Models: The Case of Social Commonsense Reasoning. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 61–72. doi:10.18653/v1/2023.trustnlp-1.7
[28] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).
[29] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. *arXiv preprint arXiv:2205.00551* (2022).
[30] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 43–53. doi:10.18653/v1/S18-2005
[31] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).
[32] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34 (2021), 2611–2624.
[33] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html
[34] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
[35] Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. *arXiv preprint arXiv:2305.11242* (2023).
[36] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *arXiv preprint arXiv:2109.03858* (2021).
[37] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437* (2018).
[38] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[39] Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour* 7, 7 (2023), 1084–1095.
[40] Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. *arXiv preprint arXiv:2301.11997* (2023).
[41] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015
[42] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 622–628. doi:10.18653/v1/N19-1063

[43] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1699–1710.

[44] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*. 1–17.

[45] Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PloS one* 11, 5 (2016), e0155036.

[46] Sourabrata Mukherjee, Atul Kr Ojha, and Ondřej Dušek. 2024. Are Large Language Models Actually Good at Text Style Transfer? *arXiv preprint arXiv:2406.05885* (2024).

[47] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[48] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133* (2020).

[49] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456* (2023).

[50] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).

[51] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2086–2105. doi:10.18653/v1/2022.findings-acl.165

[52] Barbara Plank. 2011. Domain adaptation for parsing. (2011).

[53] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5740–5745.

[54] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000* (2018).

[55] Tamara Quiroga, Felipe Bravo-Marquez, and Valentin Barriere. 2025. Adapting Bias Evaluation to Domain Contexts using Generative Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 28043–28054.

[56] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910* (2021).

[57] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[58] Evan TR Rosenman, Santiago Olivella, and Kosuke Imai. 2023. Race and ethnicity data for first, middle, and surnames. *Scientific data* 10, 1 (2023), 299.

[59] Nikil Roashan Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2022. The tail wagging the dog: Dataset construction biases of social bias benchmarks. *arXiv preprint arXiv:2210.10040* (2022).

[60] Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337* (2022).

[61] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).

[62] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. " I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. *arXiv preprint arXiv:2205.09209* (2022).

[63] Rui Song, Fausto Giunchiglia, Yingji Li, Lida Shi, and Hao Xu. 2023. Measuring and mitigating language model biases in abusive language detection. *Information Processing & Management* 60, 3 (2023), 103277.

[64] Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168* (2021).

[65] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1630–1640. doi:10.18653/v1/P19-1159

[66] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503* (2022).
[67] Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259* (2021).
[68] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219* (2023).
[69] L Monique Ward and Petal Grower. 2020. Media and the development of gender role stereotypes. *Annual Review of Developmental Psychology* 2, 1 (2020), 177–199.
[70] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics* 6 (2018), 605–617.
[71] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032* (2020).
[72] Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. *arXiv preprint arXiv:2110.00672* (2021).
[73] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.
[74] Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: Finding important examples for fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14593–14601.

A Generation Details

This section provides additional details on the selection of the in-domain examples used for prompting and on the validation of the generated templates. To ensure a fair comparison, we exclude NOEs from the prompting pool, so that the adaptation process does not rely on the same examples later used as evaluation references. We then apply additional filtering criteria to the in-domain examples. Specifically, we remove sentences with fewer than five words. For EEC-derived templates, we discard candidate examples with toxicity scores above 0.5, as predicted by the `Cardiffnlp-offensive` model. For IPTTS, we preserve the original toxic/non-toxic distribution by selecting toxic examples when the source template is toxic, and non-toxic examples otherwise; when no toxic in-domain examples are available, we fall back to non-toxic ones.

Identity terms are instantiated using popular male names obtained from the U.S. Social Security Administration (SSA). After generation, we apply a validation step to ensure that each adapted template preserves the selected identity term, contains at least four words, and is not identical to the original template. When any of these conditions is not satisfied, we resample the 15 in-domain examples and prompt the model again until a valid adapted template is obtained.

B Inter-Annotator Agreement

We report inter-annotator agreement as complementary evidence for the human evaluation. For domain transfer strength, based on categorical domain judgments, Krippendorff's alpha with nominal distance is $\alpha = 0.802$. For content preservation, rated on a 1–5 Likert scale, Krippendorff's alpha is $\alpha = 0.563$ with interval distance and $\alpha = 0.703$ with tolerant interval distance, where a difference of one point is counted as no error.

These results indicate sufficient agreement among the three annotators, especially given the subjective nature of evaluating content preservation and domain style across different domains.

C Lexical similarity analysis

To provide complementary evidence for the domain-transfer results, we compared the unigram vocabulary distributions of the adapted templates against those of the original target-domain corpora using two distributional measures: cosine similarity over relative token frequencies and Jensen–Shannon divergence. Let p and q denote the normalized word-frequency vectors of an

adapted set and a target-domain corpus, respectively. Cosine similarity, $\cos(p, q) = \frac{p \cdot q}{\|p\| \|q\|}$, measures how similarly words are distributed across the two corpora, with higher values indicating greater lexical alignment. Jensen–Shannon divergence, $\mathrm{JSD}(p, q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m)$, where $m = \frac{1}{2}(p + q)$, measures the difference between the two distributions, with lower values indicating greater similarity.

As shown in Figure 12, the adapted templates are generally closest to their intended target domain under these distributional measures. This pattern is clearest for Tweets and IMDb, whose adapted templates achieve the highest cosine similarity and among the lowest Jensen–Shannon divergence values with their corresponding target corpora. In contrast, Wikipedia Talk Pages shows weaker lexical alignment, which is consistent with the lower domain-transfer scores reported in the main text and with its greater lexical distance from the source templates.

Overall, these results support the interpretation that the adaptation process shifts the lexical profile of the templates toward the target domain, while also confirming that this shift is less pronounced for Wikipedia Talk Pages than for Tweets and IMDb.

(a) Jensen Shannon

(b) Cosine similarity

Fig. 12. Corpus-level lexical similarity between adapted templates and target-domain corpora, measured with Jensen–Shannon divergence (left; lower is better) and cosine similarity over relative token frequencies (right; higher is better).