Adapting Bias Evaluation to Domain Contexts using Generative Models

Tamara Quiroga[1,2,3,4] **Felipe Bravo-Marquez**[1,3,4] **Valentin Barriere**[1,3]
[1]Department of Computer Science, University of Chile.
[2]Department of Computer Science, Pontifical Catholic University of Chile
[3]National Center for Artificial Intelligence, Chile
[4]Millennium Institute for Foundational Research on Data, Chile
t.quiroga@uc.cl {fbravo,vbarriere}@dcc.uchile.cl

Abstract

Numerous datasets have been proposed to evaluate social bias in Natural Language Processing (NLP) systems. However, assessing bias within specific application domains remains challenging, as existing approaches often face limitations in scalability and fidelity across domains. In this work, we introduce a domain-adaptive framework that utilizes prompting with Large Language Models (LLMs) to automatically transform template-based bias datasets into domain-specific variants. We apply our method to two widely used benchmarks—*Equity Evaluation Corpus* (EEC) and *Identity Phrase Templates Test Set* (IPTTS)—adapting them to the Twitter and Wikipedia Talk data. Our results show that the adapted datasets yield bias estimates more closely aligned with real-world data. These findings highlight the potential of LLM-based prompting to enhance the realism and contextual relevance of bias evaluation in NLP systems.[1]

1 Introduction and Related Work

The rapid adoption of NLP systems has been fueled by domain- and task-specific models that deliver strong performance across a wide range of applications (Beltagy et al., 2019; Gururangan et al., 2020; Nguyen et al., 2020). However, these models have also been shown to replicate social biases, raising significant ethical concerns (Blodgett et al., 2020; Abid et al., 2021; Gallegos et al., 2024).

In response, numerous datasets have been developed to assess social bias, typically through either manually designed templates (Nozza et al., 2021; Kirk et al., 2021; Smith et al., 2022) or naturally occurring examples (NOEs) extracted from corpora (Dhamala et al., 2021; Levy et al., 2021).

Template-based datasets, composed of controlled sentence, offer scalability and have enabled systematic evaluations across various attributes and languages (Mei et al., 2023; Costa-jussà et al., 2023). However, their synthetic nature often results in sentences that diverge from real-world language use (Levy et al., 2021; Seshadri et al., 2022). By contrast, NOEs reflect authentic language patterns drawn from real-world domains such as Wikipedia (Webster et al., 2018), Twitter (Naous et al., 2024; Barriere and Cifuentes, 2024b), and Reddit (Barikeri et al., 2021). While more realistic, NOEs are costly to collect and annotate, with feasibility varying across domains and demographic groups.

While both strategies have significantly advanced bias evaluation, neither approach offers a scalable and adaptable solution for evaluating bias across diverse domains. This is a pressing need, as model behavior shifts across application domains (Hendrycks et al., 2020).

In this work, we propose a novel approach to generate domain-adapted bias measurement datasets by transforming template-based benchmarks into versions tailored to specific domains. Our method leverages prompting techniques to provide a simple, automated, and cost-effective solution that harnesses the power of LLMs—particularly in text style transfer tasks (Reif et al., 2022; Suzgun et al., 2022; Luo et al., 2023)—and is applicable across domains, as it does not rely on any domain-specific features.

We focus on social bias defined as systematic disparities in model behavior across demographic groups, and validate our approach with sentiment analysis and toxicity detection tasks. Specifically, we adapt two widely used template datasets—EEC and IPTTS (Kiritchenko and Mohammad, 2018; Dixon et al., 2018)—to the Twitter and Wikipedia Talk (WT) domains, and compare bias estimates from the adapted templates against those derived from NOEs.

Our contributions are threefold: *(i)* we propose a domain-adaptive framework for transforming

[1] Our code and data are available online.

template-based bias datasets using LLM prompting; *(ii)* we introduce an evaluation method to assess bias measurement effectiveness across social groups; and *(iii)* we provide empirical evidence that our adapted templates better align with the Twitter and WT domains.

2 Method

Figure 1 provides an overview of our approach, while Figure 3 illustrates how we evaluate its effectiveness.

2.1 Terminology

Following Czarnowska et al. (2021), we define a **sensitive attribute** $\mathcal{S}$ (e.g., gender) as a category potentially subject to bias. Each $\mathcal{S}$ is represented by a set of **protected groups** $\mathcal{G}$ (e.g., $\{$*female,male*$\}$), where each group $G \in \mathcal{G}$ is associated with a set of **identity terms** I_G (e.g., $\{$*she,woman*$\}$)

Let M denote the model under evaluation, $\mathcal{D}$ the application domain, and $B_{M,\mathcal{S},\mathcal{D}}$ the bias exhibited by M in domain $\mathcal{D}$ with respect to attribute $\mathcal{S}$.

We define $T = \{t^1, \ldots, t^N\}$ as a set of templates used to measure bias related to $\mathcal{S}$, by substituting identity terms I_G corresponding to each group $G \in \mathcal{G}$.

2.2 Template Adaptation

We adapt the template set T to match the linguistic style of the target domain $\mathcal{D}$, enabling more context-aware bias estimation $B_{M,\mathcal{S},\mathcal{D}}$. Each template $t^i \in T$ is rewritten by a generator $g_{\mathcal{D}}$ to produce a domain-adapted version $t^i_{\mathcal{D}} = g_{\mathcal{D}}(t^i)$, yielding the adapted set $T_{\mathcal{D}} = \{t^1_{\mathcal{D}}, \ldots, t^N_{\mathcal{D}}\}$. Figure 1 illustrates this transformation for a single template.

We denote the original and adapted domains as $\mathcal{D}_T$ and $\mathcal{D}_{T_{\mathcal{D}}}$, with corresponding bias estimates $B_{M,\mathcal{S},\mathcal{D}_T}$ and $B_{M,\mathcal{S},\mathcal{D}_{T_{\mathcal{D}}}}$.

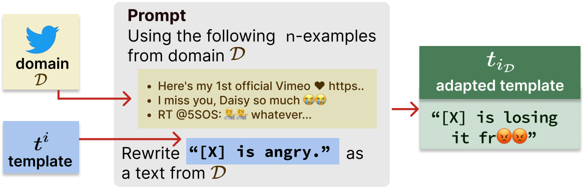

domain
$\mathcal{D}$
t^i
template
Prompt
Using the following n-examples from domain $\mathcal{D}$
• Here's my 1st official Vimeo ♥ https..
• I miss you, Daisy so much 😭😭
• RT @5SOS: 👯👯 whatever...
Rewrite “[X] is angry.” as a text from $\mathcal{D}$
$t_{i_{\mathcal{D}}}$
adapted template
“[X] is losing it fr😡😡”

Figure 1: Illustration of the template adaptation process, where an LLM rewrites a template to match the linguistic style of the target domain.

2.3 Evaluation

Our goal is to evaluate whether the set $T_{\mathcal{D}}$ better captures bias $B_{M,\mathcal{S},\mathcal{D}}$ than the original set T.

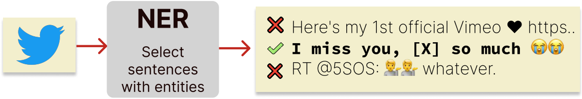

NER
Select sentences with entities
✗ Here's my 1st official Vimeo ♥ https..
✓ I miss you, [X] so much 😭😭
✗ RT @5SOS: 👯👯 whatever.

Figure 2: Illustration of the procedure used to extract NOEs from the target domain using NER.

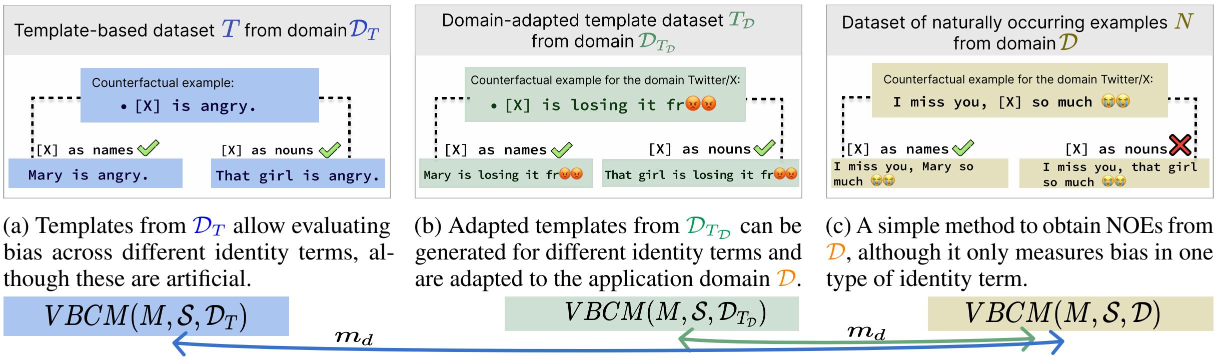
Bias Estimation in the Application Domain Estimating $B_{M,\mathcal{S},\mathcal{D}}$ can be costly, often requiring domain-specific bias datasets involving API expenses or manual filtering (Barikeri et al., 2021; Webster et al., 2018). To reduce these costs, we limit estimation to cases where $\mathcal{S}$ can be approximated using named entities as proxies for social groups—a strategy used in prior work (Barriere and Cifuentes, 2024b; Naous et al., 2024).

Following this approach, we generate NOEs by identifying instances in $\mathcal{D}$ with named entities, using Named Entity Recognition (NER), yielding a dataset $N = \{n^1, \ldots, n^{|N|}\}$. The process is illustrated in Figure 2.

While efficient, this method depends on the presence and quality of named entities and cannot measure bias beyond entity-based definitions of $\mathcal{S}$, since I_G can be only define as entities. Nonetheless, it offers a useful estimate of $B_{M,\mathcal{S},\mathcal{D}}$ for evaluating our approach.

Vector Background Comparison Metric We estimate bias using the Vector Background Comparison Metric (VBCM) introduced by Czarnowska et al. (2021). VBCM represents disparities across protected groups as a vector whose components capture the deviation of each group $G_i \in \mathcal{G}$ from a reference background β. It relies on a scoring function ϕ, which assigns scores to example subsets, and a distance function d, which quantifies discrepancies between score distributions. Our bias estimator of $B_{M,\mathcal{S},\mathcal{D}}$ is denoted as $\mathrm{VBCM}(M, \mathcal{S}, \mathcal{D})$. The formal definition and a worked example of this metric are provided in Appendix A.2.

Bias Comparison We assess if LLM-adapted templates $T_{\mathcal{D}}$ better reflect real-world biases in $\mathcal{D}$ by comparing their VBCM scores against those from original templates T, using N as reference. To quantify this, we define a distance metric m_d to compare bias vectors across domains. This comparison is illustrated in Figure 3.


Template-based dataset T from domain $\mathcal{D}_T$
Counterfactual example:
• [X] is angry.
[X] as names ✓
[X] as nouns ✓
Mary is angry.
That girl is angry.
Domain-adapted template dataset $T_\mathcal{D}$ from domain $\mathcal{D}_{T_\mathcal{D}}$
Counterfactual example for the domain Twitter/X:
• [X] is losing it fr😡😡
[X] as names ✓
[X] as nouns ✓
Mary is losing it fr😡😡
That girl is losing it fr😡😡
Dataset of naturally occurring examples N from domain $\mathcal{D}$
Counterfactual example for the domain Twitter/X:
I miss you, [X] so much 😭😭
[X] as names ✓
[X] as nouns ✗
I miss you, Mary so much 😭😭
I miss you, that girl so much 😭😭
$VBCM(M, \mathcal{S}, \mathcal{D}_T)$
$VBCM(M, \mathcal{S}, \mathcal{D}_{T_\mathcal{D}})$
$VBCM(M, \mathcal{S}, \mathcal{D})$
m_d
m_d

(a) Templates from $\mathcal{D}_T$ allow evaluating bias across different identity terms, although these are artificial.

(b) Adapted templates from $\mathcal{D}_{T_\mathcal{D}}$ can be generated for different identity terms and are adapted to the application domain $\mathcal{D}$.

(c) A simple method to obtain NOEs from $\mathcal{D}$, although it only measures bias in one type of identity term.

(d) Compare biases measured in a model M with respect to a sensitive attribute $\mathcal{S}$.

Figure 3: Comparison of strategies for generating counterfactuals from $\mathcal{D}_T$ to $\mathcal{D}_{T_\mathcal{D}}$ to approximate $\mathcal{D}$.

3 Experimental Setup

3.1 Bias

We evaluate social bias for two widely studied sensitive attributes $\mathcal{S}$: **nationality** (Ahn and Oh, 2021; Venkit et al., 2023) and **gender** (Kurita et al., 2019; Parrish et al., 2022; Kotek et al., 2023).

These attributes were chosen for their complementary representational properties: nationality admits a broad set of groups, enabling fine-grained comparison of bias vectors, while gender can be expressed via both proper names and common nouns, allowing a richer analysis of lexical variation.

Nationality bias We consider 38 countries, each represented by 50 popular personal names, following Barriere and Cifuentes (2024a,b).

Gender bias We define four groups: *female-names*, *female-nouns*, *male-names*, and *male-nouns*. The name groups comprise the top 50 names with at least 75% gender association in the U.S. SSA database, and the noun groups consist of 11 terms per gender sourced from Meade et al. (2022) (Appendix A.1).

VBCM Following Zayed et al. (2023), VBCM uses ϕ as the positive output probability and d as demographic parity $DP(x, y) = 1 - |x - y|$. The background is the average across groups following Levy et al. (2023). We compare VBCMs using MAE and Pearson correlation; lower MAE and higher correlation indicate consistent bias measurements across template sets.

3.2 Templates studied

We apply our methodology to EEC and IPTTS datasets, both originally created in English. EEC, introduced by Kiritchenko and Mohammad (2018), targets gender and racial bias in emotion regression and has inspired multilingual bias evaluations (Câmara et al., 2022). IPTTS, proposed by Dixon et al. (2018), is a synthetic dataset for measuring biases in toxicity classification with balanced toxic and non-toxic templates. It has been widely used to assess and mitigate bias (Dixon et al., 2018; Park et al., 2018; Borkan et al., 2019; Zayed et al., 2023).

For our analysis of nationality and gender bias, we select instances where identity terms can be substituted with predefined names or gendered nouns. Dataset sizes are provided in Appendix A.3.

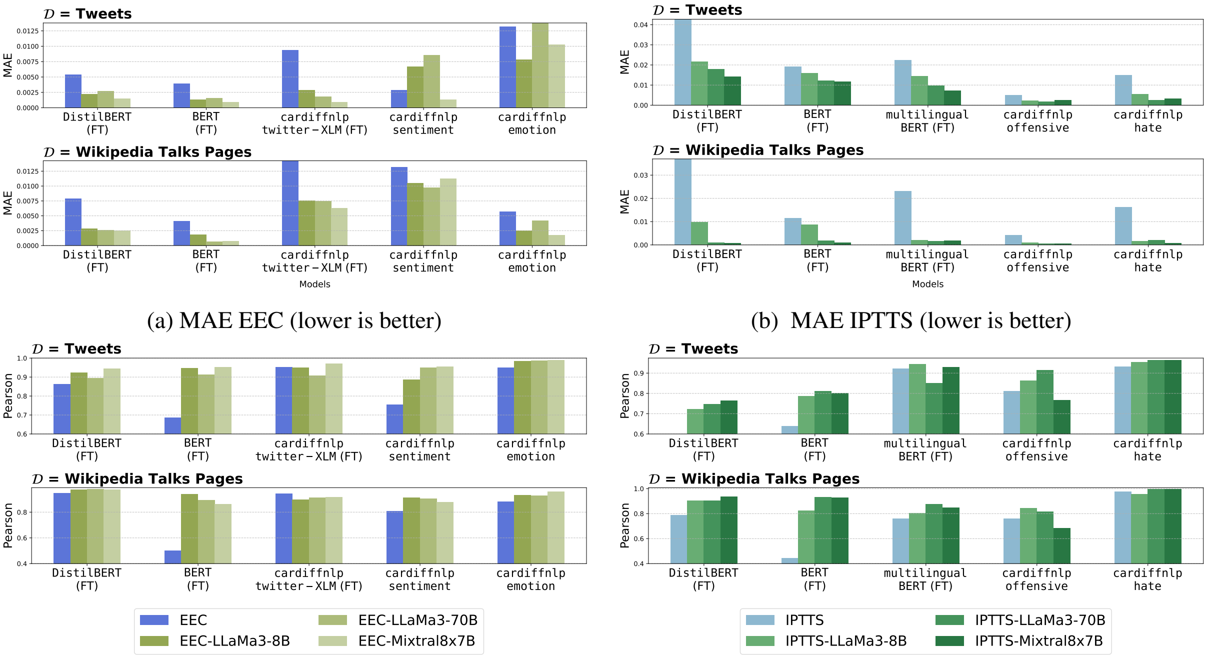
3.3 Application Domains

We focus on two real-world domains, $\mathcal{D}$: Twitter and Wikipedia Talk Pages. These domains were selected because prior work (Câmara et al., 2022; Park et al., 2018; Dixon et al., 2018; Zayed et al., 2023) has examined social bias in NLP systems trained on them using synthetic data.

Twitter is a microblogging platform characterized by brief, informal posts, often featuring slang and abbreviations due to its 280-character limit (Imran et al., 2016).

Wikipedia Talk Pages is a space for informal, interactive discussions among Wikipedia contributors, reflecting the dynamic and conversational nature of online communication.

Naturally Occurring Examples To construct NOE-based datasets for the Twitter and Wikipedia domains, we use the EuroTweets dataset (Mozetič et al., 2016) and the Wikipedia Talk Pages test set (Wulczyn et al., 2017), filtering instances that contain named entities using SpaCy's NER (Vasiliev, 2020). From these, only 475 of 8,851 EuroTweets and 1,101 of 20,000 Wikipedia



(a) MAE EEC (lower is better)

(b) MAE IPTTS (lower is better)

(c) Pearson EEC (higher is better)

(d) Pearson IPTTS (higher is better)

Figure 4: Metrics of VBCM scores between $\mathcal{D}$, $\mathcal{D}_T$ and $\mathcal{D}_{T_\mathcal{D}}$ across domains and models.

Talk instances qualify as containing valid NOEs. To balance the dataset, we select 388 examples from the Wikipedia set. These results indicate that even with a simple entity-based filtering method, only a small subset of the data is suitable for NOE-based bias evaluation.

3.4 Generated Templates

To generate domain-adapted templates, we prompt an LLM to rewrite a given template—evaluated with an identity term—using $n = 15$ domain-specific examples, while preserving the original meaning and emotional tone. Separate template sets are generated depending on whether the identity term corresponds to a name or a gendered noun. The full prompt is provided in Appendix A.4.

Domain examples $\mathcal{D}$ are drawn from EuroTweets and Wikipedia Talk Pages. Unlike NOEs, which require examples with specific named entities, our method can use any sentence from the domain as input; however, to ensure reliable comparison, we exclude instances that contain NOEs.

Datasets are generated using three open-source LLMs: `LLaMA3-8B`, `LLaMA3-70B` (Dubey et al., 2024), and `Mixtral-8x7B` (Jiang et al., 2024).

To evaluate content preservation, we compute cosine similarity between original and adapted templates. The results show consistent semantic alignment, as detailed in Appendix A.10.

3.5 Models and Task Evaluated

For sentiment regression, we fine-tune pretrained models on the SemEval-2018 Task 1 dataset (Mohammad et al., 2018) and include two off-the-shelf models. For toxicity classification, we fine-tune on Wikipedia Talk data and evaluate two additional pretrained models. See Appendix A.5 for details.

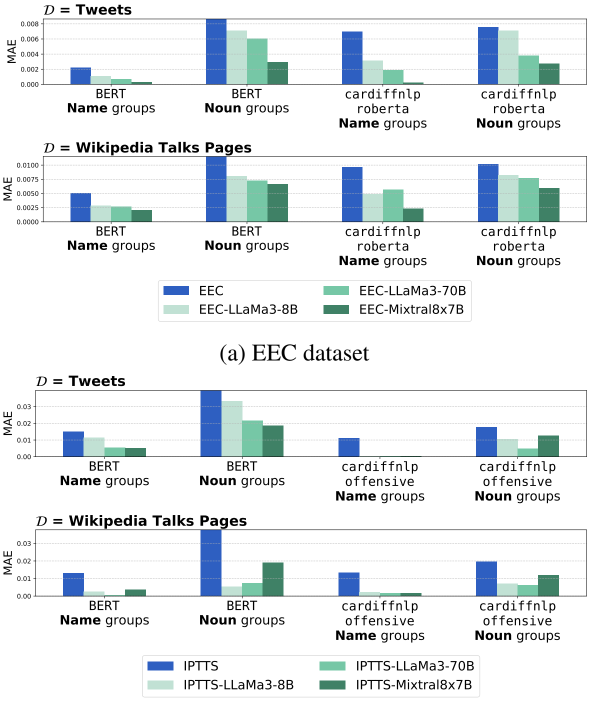
4 Results

Across domains, tasks, and models, there is a clear trend that the bias estimation on the translated templates domain $\mathcal{D}_{T_\mathcal{D}}$ is more similar in distance and in correlation to the one from application domain $\mathcal{D}$ than from the one of the original template domain $\mathcal{D}_T$, validating our hypothesis.

Nationality Bias Six adapted versions of the two template datasets were generated by combining three LLMs with the two application domains. Figure 4 show the distance and similarity between the VBCM of the real-life data and the ones calculated in the template domains $\mathcal{D}_T$ and $\mathcal{D}_{T_\mathcal{D}}$.

Across both datasets, bias estimates vary depending on the model and domain. In the worst cases, the original EEC reaches only 0.49 correlation with NOEs, while IPTTS drops to 0.24 — highlighting the limitations of relying solely on curated datasets like EEC or IPTTS for bias evaluation.

For EEC, the adapted versions consistently re-



(a) EEC dataset

(b) IPTTS dataset

Figure 5: MAE between NOEs ($\mathcal{D}$) and original ($\mathcal{D}_T$) vs LLM-adapted ($\mathcal{D}_{T_\mathcal{D}}$) in EEC and IPTTS templates in gender bias across different groups.

duce MAE (Figure 4a) and improve correlation with NOEs (Figure 4c), especially in the Twitter domain. Mixtral achieves the best performance, while LLaMA models perform similarly in most cases. A similar pattern is observed with IPTTS: the adapted versions lower MAE (Figure 4b) and increase correlation (Figure 4d). LLaMA models slightly outperform Mixtral. Notably, both LLMs generate appropriate samples despite IPTTS's sensitive content.

In summary, LLM-adapted datasets ($\mathcal{D}_{T_\mathcal{D}}$) align more closely with NOEs ($\mathcal{D}$) bias estimates than usual templates ($\mathcal{D}_T$). Mixtral performs best on EEC, while LLaMA excels on IPTTS—likely reflecting differences in dataset characteristics, particularly the presence of toxicity, which poses distinct challenges for generative models in bias evaluation.

Gender Bias Figure 5 presents MAE scores for the VBCM metric, comparing bias estimates from naturally occurring examples (NOEs), original templates ($\mathcal{D}_T$), and domain-adapted templates ($\mathcal{D}_{T_\mathcal{D}}$) for both EEC and IPTTS datasets. Results are shown for two types of identity terms—names and gendered nouns—and across two domains: Tweets and Wikipedia Talk Pages.

Across all configurations, domain-adapted templates ($\mathcal{D}_{T_\mathcal{D}}$) consistently yield lower MAE scores with respect to NOEs compared to the original templates, for both names and noun-based identity terms. This demonstrates the adaptability of our approach to generate context-sensitive templates tailored to different forms of identity representation, and highlights the importance of aligning template style with the linguistic characteristics of the target domain.

5 Conclusion and Future Work

We introduced a domain-adaptive framework for bias evaluation that uses LLM prompting to connect template-based datasets with domain-specific text. The approach is automated, scalable, cost-effective, and domain-agnostic, making it applicable across diverse settings.

Across sentiment and toxicity tasks, and for both nationality and gender, our experiments on EEC and IPTTS show that the adapted templates yield bias estimates more closely aligned with naturally occurring examples, using both fine-tuned and off-the-shelf models. These results suggest that LLMs can improve the realism and adaptability of bias evaluation, addressing a key limitation of existing methods.

Future work will extend the framework to additional datasets and model families, and will rigorously test whether LLM-generated adaptations preserve intended semantics without introducing new biases. Because our comparisons span different domains, human evaluation is essential to validate content preservation and ensure the substantive meaning of fairness assessments.

6 Limitations

While our work advances the evaluation of social biases in domain-specific contexts, several limitations remain. First, our analysis relies on counterfactual examples that use named entities to represent social groups. This design supports controlled experimentation, but using names or gendered nouns as proxies—common in prior work—can oversimplify or misrepresent complex social categories.

Second, although prompting LLMs with domain-specific examples enables efficient generation of adapted templates, the process is susceptible to semantic drift and hallucinations. Our similarity analyses indicate reasonable alignment between original and adapted templates, but they cannot fully rule out unintended shifts in meaning.

Third, our gender analyses adopt a binary framework, excluding non-binary and gender-diverse identities, which limits the scope and inclusivity of our findings.

7 Ethics Statement

Measuring social bias is inherently complex, and no single approach can guarantee comprehensive detection or mitigation of harmful patterns. Our domain-adapted datasets are intended to enhance realism and contextual relevance, but they should be regarded as diagnostic tools rather than definitive indicators of fairness. Since the datasets rely on LLM-generated content, their use requires careful review and validation to mitigate risks such as hallucinations or inconsistencies.

Acknowledgments

This work was partially financed with the grant U-INICIA 2024 from the Vicerrectoría de Investigación y Desarrollo (VID) number UI-011/24 "Estudios de sesgos sociales en modelos de lenguajes largos", ANID Chile National Doctoral Scholarship 21220586 (TQ), the grant National Center for Artificial Intelligence CENIA FB210017 Basal ANID, and ANID Chile Millennium Science Initiative Program IMFD ICN17_002.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955.

Valentin Barriere and Sebastian Cifuentes. 2024a. 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)*, pages 1511–1518.

Valentin Barriere and Sebastian Cifuentes. 2024b. 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*, pages 569–579.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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)*, pages 3615–3620.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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*, pages 491–500.

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*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

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*.

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*, 9:1249–1267.

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*, pages 862–872.

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*, pages 67–73.

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*.

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*, pages 1–79.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8342–8360. Association for Computational Linguistics (ACL).

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

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*.

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*.

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*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

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*, 34:2611–2624.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480.

Sharon Levy, Neha 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898.

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*, pages 1699–1710.

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*, pages 1–17.

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

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

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*, pages 2086–2105.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848.

Alejandro Salinas, Amit Haim, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. In *NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning (TSRML)*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023*, pages 116–122. Association for Computational Linguistics (ACL).

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*.

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*, 6:605.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532.

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*, pages 1391–1399.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the name of fairness: assessing the bias in clinical record de-identification. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 123–137.

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*, volume 37, pages 14593–14601.

A Appendix

A.1 Identity terms for noun-groups of gender bias

Male-noun = { *that man, that boy, that actor, that waiter, my dad, my brother, my father, my husband, my son, my uncle, my grandfather* }

Female-noun = { *that woman, that girl, that actress, that waitress, my mom, my sister, my mother, my wife, my daughter, my aunt, my grandmother* }

A.2 VBCM

A.2.1 Definition

To compare biases, we adopt the Vector Background Comparison Metric (VBCM) introduced by Czarnowska et al. (2021). VBCM represents bias across protected groups as a vector whose i-th component measures the disparity between the score of group G_i and that of a designated reference group—called the *background* β.

Formally let $S' = \{S'_1, \dots, S'_{|S|}\}$ be a collection of evaluation example sets, where each S'_j corresponds to all perturbations of template t_j when evaluated with identity terms from each group. Let $S'^{G_i}_j$ denote the subset of perturbations associated with a single protected group $G_i \in \mathcal{G}$. For example, if $t_1 = \{$`[MASK]` is angry.$\}$ and the identity terms in group g are $\{$*Mary*, *Kate*, *Emma*$\}$, then $S'^{G}_1 =$ $\{$Mary is angry., Kate is angry., Emma is angry.$\}$. The VBCM for a model M, a set of protected groups $\mathcal{G}$, and a set of templates T, is then computed as follows:

$$\left\{ \frac{1}{|S'|} \sum_{S'_j \in S} d\left(\phi(\beta^S), \phi(S_j'^{g_i})\right) \right\}_{G_i \in \mathcal{G}} \tag{1}$$

where β^S denotes the background group, ϕ represents a scoring function that computes the score for a subset of examples, and d is a comparison function that quantifies the discrepancy between the set of scores of examples.

Since VBCM yields a bias vector for each template set T, as well as for its variants $T_{\mathcal{D}}$ and N, we assess their similarity using both correlation and distance-based measures. High correlation and low distance between vectors indicate that different template sets reveal consistent bias measures.

A.2.2 Example for calculation of VBCM

To illustrate the concepts discussed in Section 2.3, we present a small example using *nationality* as the sensitive attribute. Three groups are considered—**Spain**, **USA**, and **France**—each associated with three representative names. Table 1 shows sample templates and their corresponding model scores.

From these values, we first compute the *background* score as the average of all template scores across countries:

$$\beta = \frac{1}{9} \sum_{\text{country, name}} \text{score} = 0.199.$$

Fairness is then assessed with the Demographic Parity (DP) metric,

$$\text{DP} = 1 - \big|\text{group\_average} - \text{Background}\big|.$$

The DP scores by country are:

$$\begin{aligned} \text{Spain:} \quad & 1 - |0.207 - 0.199| = 0.992, \\ \text{USA:} \quad & 1 - |0.193 - 0.199| = 0.994, \\ \text{France:} \quad & 1 - |0.196 - 0.199| = 0.997. \end{aligned}$$

Finally, the Vector Background Comparison Metric (VBCM) is expressed as

$$\text{VBCM} = (0.992,\ 0.994,\ 0.997).$$

This example demonstrates how VBCM captures disparities between protected groups relative to the overall background.

A.3 Dataset sizes for NOEs and Templated based dataset used

Dataset	Template	Pertubations per		
		Country (names)	Gender (names)	Gender (nouns)
EEC	144	7200	7200	1440
IPTTS	864	43200	43200	8640
NOEs Twitter	475	23750	23750	-
NOEs Wikipedia Talks	388	19400	19400	-

Table 2: Dataset sizes for template-based (EEC, IPTTS) and NOE corpora—reporting templates and total perturbations per attribute; "–" denotes not available.

A.4 Prompt

In Figure 6 we show the prompt use for adapted templates from EEC and IPTTS por gender and nationality bias.

For each template $t_i \in T$ we evaluate the template in using a identity term and we ask the model

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 semantics 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 rewrite text, without any
introduction or explanation.
```

Figure 6: Prompt

In the case where identity terms are represented by names for counterfactual evaluation, we consider that generative models can reflect name-related biases (Kirk et al., 2021; Salinas et al., 2024). To mitigate this, we use the 200 most common U.S. male names, which have been shown to carry fewer negative biases and to perform better in tasks such as text generation and entity recognition (Wolfe and Caliskan, 2021; Xiao et al., 2023; Wan et al., 2023).

Table 1: Example templates evaluated by nationality ($T = \{\text{Spain}, \text{USA}, \text{France}\}$) with $|I| = 3$ names per group.

Template	Spain (score)	USA (score)	France (score)
I like {person}.	I like María (0.20) I like José (0.25) I like Carmen (0.17)	I like Emily (0.18) I like John (0.21) I like Michael (0.19)	I like Chloé (0.19) I like Lucas (0.22) I like Emma (0.18)

A.5 Models and Tasks Evaluated

Sentiment regression. We fine-tune `BERT`, `DistilBERT`, and `cardiffnlp/twitter-roberta-base` on the SemEval-2018 Task 1 dataset (Mohammad et al., 2018). We also include two off-the-shelf models—`cardiffnlp/twitter-roberta-base-emotion` and `cardiffnlp/twitter-xlm-roberta-base-sentiment` —adapted to regression via linear aggregation of class outputs.

Toxicity classification. We fine-tune `BERT`, `DistilBERT`, and `bert-base-multilingual-cased` on the Wikipedia Talk corpus (Dixon et al., 2018), and evaluate two additional pretrained classifiers: `cardiffnlp/twitter-roberta-base-hate` and `cardiffnlp/twitter-roberta-base-offensive`.

A.6 Hardware

Our experiments were conducted on a high-performance computing cluster, using nodes equipped with multiple NVIDIA RTX 6000 GPUs.

A.7 Implementation Details

We implement our models using PyTorch and leverage HuggingFace libraries for both pretrained and off-the-shelf models, including BERT, DistilBERT, multilingual BERT, and the RoBERTa model from cardiffnlp. We fine-tuned the models using the AdamW optimizer with a learning rate of 5e-5 and trained for 3 epochs with a batch size of 16. A linear learning rate scheduler was applied with no warm-up steps, and the total number of training steps was set according to the dataset size.

For generative modeling, we use LLaMA3-8B[2], LLaMA3-70B[3], and Mixtral 8x7B[4] via HuggingFace. We apply low temperature values (0 and 0.01) during generation, as we observed minimal variation in output quality between these settings. The maximum generation length is set to 102 tokens for the Wikipedia domain and 40 tokens for the Twitter domain.

A.8 Usage Intent and Licensing of Datasets

We used datasets and models that are publicly available and distributed under licenses that permit research use. Specifically, the datasets (e.g., EEC, IPTTS, EuroTweets, Wikipedia Talk) were obtained from official repositories or associated GitHub pages linked in their original publications, which explicitly allow academic usage.

Similarly, the pretrained and off-the-shelf models used in our work (e.g., BERT, distilBERT, multilingualBERT, LLaMA, Mixtral) were accessed through the HuggingFace platform, which provides detailed licensing information for each model. All models used were distributed under licenses that allow non-commercial and research applications.

In all cases, we adhered to the terms of use provided by the original authors and platforms, and we properly cited the relevant papers to acknowledge their work.

A.9 Comparison of Template Characteristics

In this section, we analyze how LLM-adapted template datasets compare to NOEs in terms of text similarity.

Figure 7 compares word counts and spelling error rates between the original EEC dataset, its LLM-adapted Twitter versions, and real tweets. The adapted versions tend to produce longer sentences, closer to tweet lengths, and exhibit more frequent and variable spelling errors, reflecting the informal nature of Twitter language.

A similar comparison for the Wikipedia Talk domain is presented in Figure 8, using the IPTTS dataset and its adapted versions alongside real comments. As with the Twitter domain, the adapted templates in this case also adjust word count and spelling patterns, bringing them closer to the characteristics observed in real data.

However, in this case it seems that LLMs have a lower performance that in Twitter domains, this

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[4] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

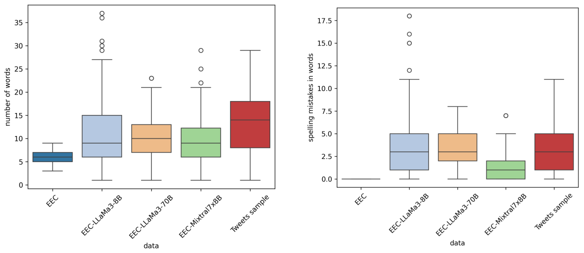
may be explains by the fact that Twittter is a domains more popular and probablt the LLMs have a better understand of how to rewrite templates. Even when the prompt do not mention the name of source of the data.

Table 3 reports the average frequency of Twitter-specific elements across the original EEC, its adapted versions, and real tweets. As expected, the original EEC contains no emojis, hashtags, mentions, or links, reflecting its controlled design. In contrast, the adapted datasets incorporate these features to varying degrees, better matching Twitter's style.

Among them, EEC-LLaMa3-8B shows the highest rates of emojis and hashtags, often exceeding real tweet counts, while EEC-LLaMa3-70B produces more moderate levels. EEC-Mixtral7x8B achieves the closest alignment to real tweets, particularly in emojis and mentions, underscoring differences in domain adaptation across models.

Template	Avg. Emoji	Avg. Hashtag	Avg. User Mention	Avg. Link
Tweets sample	0.37	0.41	0.85	0.35
EEC	0	0	0	0
EEC-LLaMa3-8B	6.20	2.31	0.24	0.21
EEC-LLaMa3-70B	3.08	0.80	0.41	0.32
EEC-Mixtral7x8B	0.38	0.15	0.15	0.01

Table 3: Average frequency of Twitter-specific elements per sentence in manually created and LLM-adapted Twitter datasets.



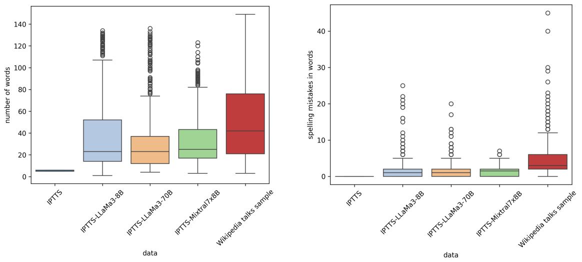
(a) Number of Words per Sentence

(b) Spelling Mistakes per Sentence

Figure 7: Comparison of word counts and spelling mistakes between EEC, adapted EEC, and Tweets samples.

A.10 Semantic Similarity Analysis

To examine whether the adapted templates preserve the meaning of the original sentences, we computed cosine similarity between the sentence embeddings of each original template and its domain-adapted counterpart. Embeddings were obtained using the `all-MiniLM-L6-v2` model from `SentenceTransformers`.



(a) Number of Words per Sentence

(b) Spelling Mistakes per Sentence

Figure 8: Comparison of word counts and spelling mistakes between IPTTS, adapted IPTTS, and Wikipedia Talks samples.

Table 4: Average cosine similarity between original and domain-adapted templates for EEC and IPTTS datasets across domains and LLMs.

Dataset	Domain	LLM	Avg. Cosine
EEC	Tweets	LLaMA3-70B	0.514
		LLaMA3-8B	0.588
		Mixtral-8x7B	0.598
	WT	LLaMA3-70B	0.651
		LLaMA3-8B	0.679
		Mixtral-8x7B	0.607
IPTTS	Tweets	LLaMA3-70B	0.606
		LLaMA3-8B	0.665
		Mixtral-8x7B	0.658
	WT	LLaMA3-70B	0.675
		LLaMA3-8B	0.717
		Mixtral-8x7B	0.701

Table 4 reports the average cosine similarity scores for both EEC and IPTTS datasets across the Twitter and Wikipedia Talk domains and for each LLM variant. Overall, the results show moderate to high similarity, indicating that the LLM-based adaptation preserves most of the original semantic content while allowing stylistic changes specific to each domain. It is worth noting that these values may slightly underestimate the true level of semantic preservation, since the embeddings are computed across different domains and the performance of the embedding model can be affected by domain mismatch.

A.11 Personally Identifying Info Or Offensive Content in Datasets

The Wikipedia Talk dataset and EuroTweets may contain personally identifying information or offensive content, as they are based on user-generated text. We did not apply additional filtering to remove such content in some of our experiments, since our objective was to assess model behavior in realistic, in-the-wild conditions where such cues may naturally occur. However, we used automated named entity recognition (NER) to select instances containing named entities, which may include personal names, without manually verifying or anonymizing them.

We acknowledge the potential risks associated with this choice and emphasize that our work does not attempt to identify individuals. Moreover, we do not release the adapted templates generated in our experiments to avoid unintended disclosure or propagation of sensitive or offensive content. All experiments were conducted in compliance with ethical research standards.

A.12 Use of AI Assistants

We occasionally used ChatGPT to assist with coding tasks, including searching for specific commands and generating functions, which were subsequently tested to ensure they behaved as expected. Additionally, AI tools were employed to support rephrasing and correct grammatical errors throughout the text.