A Novel Metric-Based Counterfactual Data Augmentation with Self-Imitation Reinforcement Learning (SIL)

K. C. Sreedhar[1], T. Kavya[2], J. V. S. Rajendra Prasad[3], V. Varshini[4]
Associate Professor, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, India[1]
Student, Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, India[2, 3, 4]

***Abstract*—The inherent biases present in language models often lead to discriminatory predictions based on demographic attributes. Fairness in NLP refers to the goal of ensuring that language models and other NLP systems do not produce biased or discriminatory outputs that could negatively affect individuals or groups. Bias in NLP models often arises from training data that reflects societal stereotypes or imbalances. Robustness in NLP refers to the ability of a model to maintain performance when faced with noisy, adversarial, or out-of-distribution data. A robust NLP model should handle variations in input effectively without failing or producing inaccurate results. The proposed approach employs a novel metric called CFRE (Context-Sensitive Fairness and Robustness Evaluation) designed to measure both fairness and robustness of an NLP model under different contextual shifts. Next, it projected the benefits of this metric in terms of experimental parameters. Next, the work integrated counterfactual data augmentation with help of Self-Imitation Reinforcement Learning (SIL) to reinforce successful counterfactual generation by enabling the model to learn from its own high-reward experiences, fostering a more balanced understanding of language. The integration of SIL allows for efficient exploration of the action space, guiding the model to consistently produce unbiased outputs across different contexts. The proposed approach demonstrates the effectiveness of our method through extensive experimentation and compared the results of the proposed metric with that of WEAT and SMART testing, and showed a significant reduction in bias without compromising the model's overall performance. This framework not only addresses bias in existing models but also contributes to a more robust methodology for training fairer NLP systems. Both the proposed metric and SIL showed better results in experimental parameters.**

Keywords*—Natural language processing; fairness, robustness; Word Embedding Association Test (WEAT); SMART testing

I. Introduction

Natural Language Processing (NLP) serves as a linchpin in enabling seamless human-computer interaction, fostering intuitive communication through interfaces like voice assistants and chatbots. It empowers the automation of text analysis, expediting tasks such as sentiment assessment, document summarization, and content categorization with unparalleled efficiency. By transcending linguistic barriers, NLP promotes global interconnectivity, facilitating multilingual translation and cultural localization.

Its contributions to AI advancements are transformative, powering sophisticated systems like personalized virtual assistants and predictive analytics. NLP is instrumental in extracting actionable insights from unstructured textual data, supporting informed decision-making in critical domains like healthcare, finance, and governance. Furthermore, it champions inclusivity by fostering the development of assistive technologies, such as speech-to-text systems and screen readers, to accommodate individuals with disabilities.

By addressing linguistic diversity and automating complex textual processes, NLP is not merely a technological tool but a catalyst for innovation and inclusivity in the digital age.

Natural Language Processing, a subfield of Artificial Intelligence, has become pivotal in automating and enhancing communication, yet its deployment raises pressing concerns around fairness and robustness. At its core, fairness in NLP pertains to the equitable and unbiased performance of language models across diverse demographic and linguistic groups. Robustness, conversely, measures a model's resilience to adversarial inputs, distributional shifts, or unexpected variations in data. Together, these dimensions are critical to ensuring the ethical and reliable use of NLP technologies.

One of the primary fairness challenges arises from biased training datasets, which reflect historical inequities, stereotypes, or regional disparities. These biases, embedded in language corpora, can perpetuate societal injustices when reflected in model outputs. For instance, gendered pronoun resolution systems may reinforce occupational stereotypes by associating women with caregiving roles and men with leadership positions.

Robustness, on the other hand, is tested when models face adversarial attacks or operate in low-resource settings. Subtle manipulations in input texts—like typos or syntax changes—can disproportionately degrade model performance. Similarly, underrepresentation of certain languages, dialects, or sociolects exacerbates the risk of exclusionary AI systems that fail to generalize effectively.

The interplay of these issues creates a dual imperative: to mitigate inherent biases while enhancing models' adaptability across varied scenarios. Ethical considerations are further compounded by the lack of standardized benchmarks for measuring fairness and robustness. Solutions often involve trade-offs, as techniques that improve robustness, like data augmentation, may inadvertently amplify biases.

Addressing these challenges requires a multi-faceted approach. Incorporating diverse, high-quality datasets and developing fairness-aware training algorithms are pivotal steps. Furthermore, interdisciplinary collaboration—spanning computational linguistics, ethics, and social sciences—can provide nuanced perspectives to inform NLP research. Regular audits, explainable AI methods, and inclusive design principles are essential to embedding trustworthiness into language technologies.

In conclusion, fairness and robustness are not merely technical hurdles but societal imperatives in the age of pervasive AI. As NLP systems permeate sensitive domains like hiring, healthcare, and legal adjudication, ensuring their ethical and equitable deployment becomes a moral obligation. The paper is organized as follows. Section I gives introduction the problem of bias in NLP. Section II gives explains types of bias in NLP. Section III gives various existing metrics for measuring bias. Section IV explains briefing, challenges of robustness and robustness contextual evaluation respectively. Experimental results is given in Section V and finally, the paper is concluded in Section VI.

1) The Problem of Bias in NLP: Bias in Natural Language Processing (NLP) refers to the systematic favoritism or prejudice exhibited by language models, often stemming from imbalances or stereotypes present in their training data. This phenomenon undermines the equity, reliability, and ethicality of NLP systems, leading to unintended discriminatory consequences. Bias is particularly critical in applications influencing high-stakes decisions, such as hiring algorithms, legal systems, and healthcare tools, where such predispositions can perpetuate societal inequities [1-3].

At its root, bias arises from the data-driven nature of NLP models, which inherit the flaws, prejudices, and imbalances embedded in the corpora used for training. When these systems process text, they often reinforce or amplify existing stereotypes, inadvertently perpetuating harm against underrepresented or marginalized groups. Addressing bias is a multifaceted challenge that requires understanding its various types and manifestations.

II. Types of Bias in NLP

1) Representation bias: This form of bias originates in training datasets that over represent certain groups or perspectives while neglecting others. For example, texts predominantly authored in English may marginalize speakers of minority languages or dialects, perpetuating cultural hegemony.

2) Stereotypical bias: Models can perpetuate harmful stereotypes, such as associating certain professions with specific genders or ethnicities. For instance, a model might predict "nurse" as a woman or "engineer" as a man based on biased correlations in training data.

3) Historical bias: Historical biases reflect long-standing societal inequities embedded in data. Even if collected neutrally, datasets often capture systemic inequalities, such as racial or gender disparities, which are then reflected in the model's predictions.

4) Selection bias: This bias arises from skewed data collection processes. If a training dataset is predominantly drawn from urban populations, for instance, the resulting model may fail to generalize to rural or less technologically advanced contexts.

5) Aggregation bias: When data from diverse groups are aggregated into a single dataset, the unique characteristics of minority groups may be overshadowed by majority trends, leading to homogenized outputs that overlook nuanced needs.

6) Interaction bias: This bias emerges during user interaction with NLP systems. For example, users' queries can introduce biases that models then propagate, such as autocomplete suggestions that reinforce prejudiced or inappropriate language.

7) Temporal bias: Temporal bias stems from the use of outdated data that fails to account for societal evolution. For instance, older datasets might include terms or perspectives that are now considered offensive or obsolete.

8) Implicit bias: Implicit biases are more subtle and embedded within the model's architecture, often surfacing in nuanced contexts such as sentiment analysis or content moderation, where subjective judgments are involved.

A. Metrics for Assessing Bias

Quantifying bias in NLP systems is a multifaceted task that requires metrics capable of identifying disparities, imbalances, and stereotypical tendencies. These metrics enable researchers to evaluate the degree of bias and its impact, facilitating informed strategies for mitigation. Below is an overview of commonly used metrics for measuring bias in NLP, along with their mathematical formulations:

1) Statistical Parity Difference (SPD): This metric evaluates whether the outcomes for different demographic groups are equally distributed.

$$SPD = P(Y=1|G=g1) - P(Y=1|G=g2)$$

- Y: Model outcome (e.g., positive or negative sentiment).
- GGG: Demographic group (g1, g2 represent different groups, e.g., male and female).
- A value of 0 indicates perfect fairness, while deviations suggest bias.

2) Equal Opportunity Difference (EOD): This metric focuses on the equality of true positive rates across groups, ensuring that all groups have equal chances of achieving favorable outcomes when eligible.

$$EOD = P(\acute{Y}=1|Y=1,G=g1) - P(\acute{Y}=1|Y=1,G=g2)$$

- Ý: Predicted outcome.
- Ensures fairness specifically for eligible or qualified individuals.

3) Conditional Demographic Disparity (CDD): This metric measures bias in model predictions while controlling for specific contextual variables.

$$CDD=P(\acute{Y}=1|X=x,G=g1)-P(\acute{Y}=1|X=x,G=g2)$$

- X: Contextual variables, such as input features.
- Helps identify disparities conditional on input attributes.

4) Word Embedding Association Test (WEAT): This metric quantifies bias in word embeddings by measuring the association between target words and attribute word sets.

$$WEAT = \frac{mean(s(w,A) - s(w,B))}{std(s(w,A) - s(w,B))}$$

- w: Target word.
- A, B: Two sets of attribute words (e.g., male- and female-associated words).
- s(w, A): Cosine similarity between w and words in set A.
- A high WEAT score indicates stronger associations, reflecting potential biases.

5) Bias Amplification Index (BAI): This measures the extent to which a model amplifies existing biases in data.

$$BAI = \frac{Bias\ in\ Model\ Output}{Bias\ in\ Training\ Data}$$

Ratios greater than 1 indicate that the model exacerbates bias.

6) Directional Bias Metric (DBM): This metric evaluates bias in sentence or text-level outputs by analyzing directional shifts in embeddings.

$$DBM = \frac{\sum_{i=1}^{n} \cos(\overrightarrow{e}, \overrightarrow{d})}{n}$$

$\overrightarrow{e}$: Embedding of Sentence i

$\overrightarrow{d}$: Bias direction vector

n: Total sentences

7) Mutual Information Difference (MID): This metric captures the disparity in the information shared between model predictions and sensitive attributes.

$$MID=I(\acute{Y};G=g1)-I(\acute{Y};G=g2)$$

- I: Mutual information between predictions Ý and group G.
- A high MID score reflects unequal representation of sensitive attributes in predictions.

8) KL Divergence for Demographic Representation (KLD): This measures the divergence between the distributions of outcomes for different demographic groups.

$$KLD(P||Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)}$$

- $P(i)$: Outcome distribution for group g1
- $Q(i)$: Outcome distribution for group g2
- Lower divergence values indicate better fairness.

9) Bias Direction Magnitude (BDM): This quantifies the degree of separation between different demographic groups in embedding space.

$$BDM = ||mean(\vec{e}_{g1}) - mean(\vec{e}_{g2})||$$

$\vec{e}_{g1}, \vec{e}_{g2}$: Embeddings for groups g1 and g2.

10) Token Probability Disparity (TPD): This metric measures bias in token-level predictions for specific sensitive terms.

$$TPD=P(token|G=g1)-P(token|G=g2)$$

Highlights disparities in word usage or token generation probabilities.

These metrics provide nuanced perspectives on bias in NLP systems, addressing its various dimensions, such as representation, prediction fairness, and embedding neutrality. Combining multiple metrics is essential for comprehensive evaluation, as bias often manifests in subtle and multifaceted ways.

B. Robustness in NLP

Robustness in NLP refers to the ability of a model to maintain performance when faced with noisy, adversarial, or out-of-distribution data. A robust NLP model should handle variations in input effectively without failing or producing inaccurate results.

C. Example of Robustness Challenges

1) Adversarial attacks: An NLP model trained to classify movie reviews as positive or negative might be tricked by inserting inconspicuous typos or irrelevant phrases. For example, changing "The movie was great!" to " The moovie was gr8!" should ideally still yield a positive classification.

2) Context sensitivity: An NLP system that performs well on one data distribution (e.g., news articles) may fail on another (e.g., social media text) if it's not robustly trained.

D. Robustness Improvement Techniques

1) Adversarial training: Including perturbed or adversarial examples during training so that the model learns to be resilient.

2) Augmentation with noisy data: Training on data that has been altered to include variations such as different spelling, slang, or paraphrasing helps models generalize better.

3) Balancing fairness and robustness: Improving fairness often involves altering the data or the training process to mitigate biases, which can sometimes reduce robustness if not done carefully. Conversely, making a model highly robust through general training methods may not necessarily address

inherent biases. The challenge lies in designing approaches that optimize both.

E. SMART Testing

SMART Testing is a methodological paradigm for systematically evaluating NLP systems across diverse dimensions, emphasizing their fairness, robustness, and adaptability. The acronym SMART encapsulates Sensitive attributes**,** Multiple subpopulations**,** Artifacts**,** Reasoning abilities, and Temporal changes**,** reflecting the multifaceted nature of NLP evaluation. While the framework does not have a universally fixed mathematical formulation, key metrics and equations can be used to assess these dimensions.

1) Sensitive attributes (Fairness metrics): This component assesses disparities in performance between demographic groups with respect to sensitive attributes like gender or ethnicity. A commonly used fairness metric is Statistical Parity Difference (SPD):

$$SPD=|P(\acute{Y}=1|G=g1)-P(\acute{Y}=1|G=g2)|$$

Where:

- Ý: Model's predicted outcome.
- G: Demographic groups (g1 and g2 represent different groups).

A value closer to zero denotes minimal bias.

2) Multiple subpopulations (Subgroup disparities): This dimension examines model performance across distinct subpopulations within the data. Disparities are quantified using subgroup metrics such as accuracy variance:

$$\text{Variance}=\frac{\sum_{i=1}^{n}(P_i-\mu)^2}{n}$$

Where,

- P_i is Model performance for subgroup i.
- μ is mean performance across all subgroups

A high variance indicates uneven performance among subgroups.

3) Artifacts (Sensitivity to spurious patterns): Artifacts represent unintended correlations in training data that can lead to spurious model predictions. Artifact sensitivity can be measured by comparing performance on artifact-augmented data to baseline data:

$$\text{Artifact Sensitivity} = \frac{Performance_{artificat}}{Performance_{baseline}}$$

Ratios significantly deviating from 1 suggest a susceptibility to artifacts.

4) Reasoning abilities (Cognitive robustness): This evaluates the model's logical and linguistic reasoning abilities under adversarial transformations or complex scenarios. Robustness against transformations is defined as:

$$Robustness\ Score(RS) = \frac{Post\ Transformation\ Accuracy}{Baseline\ Accuracy}$$

It is stated that higher scores signify greater resistance to input perturbations.

5) Temporal changes (Adaptability over time): This aspect assesses how well the model performs as linguistic norms evolve. Temporal robustness is evaluated by measuring performance deviation across time-stamped datasets.

$$Temporal\ Deviation(TD) = |Performance_{t1} - Performance_{t2}|$$

It is stated that smaller deviations reflect higher adaptability to temporal variations.

6) Aggregated SMART score: To provide a unified view, an aggregated score can be computed as a weighted combination of the individual dimensions:

$$\text{SMART Score}=w1\cdot SPD+w2\cdot Variance+w3\cdot Sensitivity+w4\cdot Robustness\ Score+w5\cdot Temporal\ Deviation$$

Where w1, w2, w3, w4, and w5 are weights reflecting the relative importance of each dimension.

III. Proposed Novel Metric-Context-Sensitive Fairness and Robustness (CFRE)

The proposed Context-Sensitive Fairness and Robustness Evaluation (CFRE) metric is designed to measure both fairness and robustness of an NLP model under different contextual shifts [3-9]. Below is the mathematical formulation of the proposed metric:

A. CFRE Metric Components

1) Fairness Impact Score (FIS): The Fairness Impact Score evaluates the difference in output distributions when the model is tested with original data (O_{orig}) and perturbed data (O_{pert}) across different demographic or context groups (G_i).

$$\text{FIS} = \frac{1}{|G|}\sum_{i=1}^{|G|} D_{KL}(P(O_{orig}|G_i)||P(O_{pert}|G_i))$$

Where

- D_{KL} is Kullback-Leibler (KL) divergence.
- $P(O_{orig}|G_i)$ and $P(O_{pert}|G_i)$ are probability distributions of outputs for groups G_i in original and perturbed cases respectively.
- $|G|$ is number of distinct groups being evaluated.

2) Robustness Contextual Evaluation (RCE): The robustness contextual evaluation (RCE) measures the stability of model predictions by computing the cosine similarity between output vectors from original and perturbed data (O_{orig}) and (O_{pert}) respectively.

$$RCE = \frac{1}{N}\sum_{j=1}^{N}\frac{o_{orig}^{j}.o_{pert}^{j}}{\left\|o_{orig}^{j}\right\|\left\|o_{pert}^{j}\right\|}$$

where

- N is the number of samples.
- O_{orig}^{j} and O_{pert}^{j} are the output vectors for j[th] sample in the original and perturbed data sets.

3) Combined CFRE score: The overall CFRE score can be weighted combination of the FIS and RCE to balance fairness and robustness.

$$CFRE = \alpha * FIS + \beta * RCE$$

Where α and β are weights that control importance of each component.

This formulation allows us to assess not just how fair is model across different contexts but also how consistently it performs when subject to contextual variations.

In the context of the CFRE metric, the interpretations for FIS RCE, and combined CFRE are given as below.

a) Fairness Impact Score (FIS):

- Interpretation: A higher FIS value indicates a greater divergence between the original and perturbed model outputs, suggesting that the model's fairness is more sensitive to contextual changes. This can mean the model exhibits potential biases when tested with varied input conditions, highlighting fairness issues.
- Lower FIS: Implies that the model maintains fairness across different demographic or context groups, showing resilience to contextual shifts.

b) Robustness Contextual Evaluation (RCE):

- Interpretation: This score reflects how similar the model's outputs remain under perturbations. A higher RCE value means the model is more robust, maintaining consistent behavior even when inputs are contextually modified.
- High RCE: Indicates strong robustness, where the model produces stable outputs across different contexts.
- Lower RCE: Suggests that the model's predictions are more context-dependent and can vary significantly with slight input changes.

c) Overall CFRE Value:

- Combined Score: The weighted sum of FIS and RCE allows us to evaluate both fairness and robustness together.
- High CFRE with balanced weights: Implies that the model is sensitive to contextual shifts (indicating fairness issues) but also robust in maintaining consistent outputs under certain conditions.
- Lower CFRE: Indicates that the model is more fair and robust across various tested contexts, demonstrating resilience and equitable behavior.

IV. Integrating Cfre Metric Into Self Imitation Learning (SIL)

A. Introduction to Self-Imitation Learning (SIL)

Self-Imitation Learning (SIL) is an advanced reinforcement learning technique that enables agents to learn from past experiences, even suboptimal ones, by revisiting previously successful trajectories. Unlike traditional reinforcement learning, which often prioritizes exploration or maximizing immediate reward signals, SIL leverages historical data to reinforce and improve upon earlier decisions. It is particularly effective in complex environments where exploration is expensive or risky, as it capitalizes on self-generated "expert" demonstrations to refine policy optimization. By integrating memory-based learning with reinforcement dynamics, SIL demonstrates resilience in solving tasks requiring long-term planning and precise decision-making [10-18].

B. Main Idea behind Self Imitation Learning (SIL)

At its core, Self-Imitation Learning revolves around the principle of leveraging an agent's historical successes as pseudo-demonstrations for future improvement. Unlike standard reinforcement learning paradigms, which discard suboptimal trajectories, SIL recognizes that even suboptimal actions can contain valuable information for solving complex tasks. This is particularly important in environments with sparse or delayed rewards, where the exploration of new policies might fail to yield immediate benefits.

SIL achieves this by employing a replay buffer, which stores trajectories (sequences of states, actions, and rewards) that yielded above-average returns. These trajectories are treated as guiding examples, and the agent revisits them during training to imitate its own past successes. This imitation process is formalized through a self-imitation loss function, which adjusts the policy to reproduce actions from successful trajectories.

The central innovation of SIL lies in its ability to balance exploitation and exploration dynamically. While traditional methods often face a trade-off between exploiting known strategies and exploring new possibilities, SIL introduces a mechanism where self-imitation augments learning efficiency without stifling exploration. This enables the agent to improve incrementally, even in scenarios where external rewards are scarce or noisy.

Moreover, SIL is robust to noise and imperfect demonstrations, as it does not rely on external expert input but instead generates its training data from its own interactions with the environment. This self-reliant nature makes it highly scalable and adaptable to diverse tasks, from robotics to game-playing.

In essence, SIL represents a shift from purely reward-driven learning to a hybrid framework that integrates self-guidance, allowing agents to harness the full potential of their past experiences for future success. By embracing both imitation and exploration, it achieves greater sample efficiency and stability in training, setting a new benchmark for learning in complex and uncertain domains.

C. Integrating CFRE with SIL

The CFRE metric is a performance measure designed to evaluate the trade-off between fairness and reward optimization in reinforcement learning. Integrating CFRE into Self-Imitation Learning (SIL) involves modifying the SIL framework to consider fairness explicitly during the learning process. The Algorithm 1 shows the CFRE integrated into SIL. Integrating the CFRE metric into Self-Imitation Learning (SIL) can effectively scale to real-world NLP systems operating in resource-constrained environments by prioritizing fairness and reward efficiency in model training. The approach allows selective reuse of high-reward, fairness-optimized trajectories, reducing computational overhead while maintaining equitable outcomes. By leveraging the CFRE metric's adaptability, the framework aligns with limited-resource constraints, improving both performance and inclusivity without excessive reliance on additional data or computing power. This ensures robust deployment of NLP systems in diverse, real-world scenarios.

Algorithm-1: CFRE-Integrated SIL

1. **Initialize**:
   a) Define the environment E, action space A, and state space S.
   b) Initialize SIL's policy $\pi_\theta(a|s)$, replay buffer B and reward function R(s,a).
   c) Set the CFRE threshold τ, which balances fairness and efficiency.
2. **Collect Experience**:
   a) Interact with the environment to generate trajectories $\tau = (s_t, a_t, r_t, s_{t+1})$ using the current policy π_θ.
   b) Add the trajectories to the replay buffer B.
3. **Compute CFRE Metric**:
   a. For each trajectory τ, compute the FIS and CRE :
   b. $CFRE(\tau)=\alpha \cdot FIS(\tau)+\beta \cdot CRE(\tau)$ Where:
      i. α,β: weights balancing fairness and reward efficiency.
      ii. $CRE(\tau)$=Sum of rewards/Length of Trajectory
      iii. FIS (τ): Fairness computed using sensitive attributes or group-specific metrics.
   c. Retain trajectories with $CFRE(\tau) \geq \tau$ in B.
4. **Update Policy:**
   Use the retained trajectories from B to compute the SIL loss:
   a) SIL loss:
   $L_{SIL} = -\log(\pi_\theta(a|s)).(R_{expected}(s) - R_{observed}(s))$
   b) Apply gradient descent to minimize L_{SIL}.
5. **Test Policy:**
   a) Evaluate the updated policy using CFRE and track performance metrics such as fairness scores, reward efficiency, and overall task accuracy.
6. **Repeat**:
   Continue the process for a predefined number of episodes or until convergence by repeating steps 2 to 5.

V. Experimental Results

The experiment was conducted using Crow-S pairs data set on Google Colab platform of python version 3.11.8. The Crow-S pairs dataset is a benchmark specifically designed to measure biases in NLP models, focusing on sensitive social attributes like gender, race, and socioeconomic status.

It consists of sentence pairs where one sentence carries subtle bias while the other is neutral, enabling the evaluation of a model's fairness by observing its scoring discrepancies. By systematically exposing latent stereotypes or prejudiced behavior in model outputs, the dataset also tests the robustness of NLP systems against biased linguistic patterns, helping to create more equitable language technologies.

At first, we project the graph showing comparison of original and perturbed scores using CFRE as given in Fig. 1. Next, we project the density over scores of WEAT, SMART testing as given in Fig. 2. Next, we project mean scores for various metrics as given in Fig. 3.Finally, we project graphs for average loss versus epochs and average reward versus epochs as given in Fig. 4. Fig. 1 to Fig. 3 showed significant improvement in results in proposed CFRE metric.

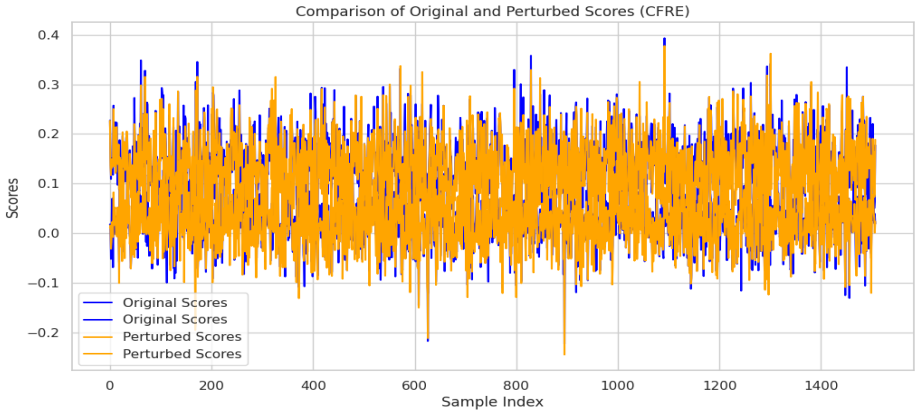

Fig. 1. Comparision of original and perturbed scores for CFRE metric.

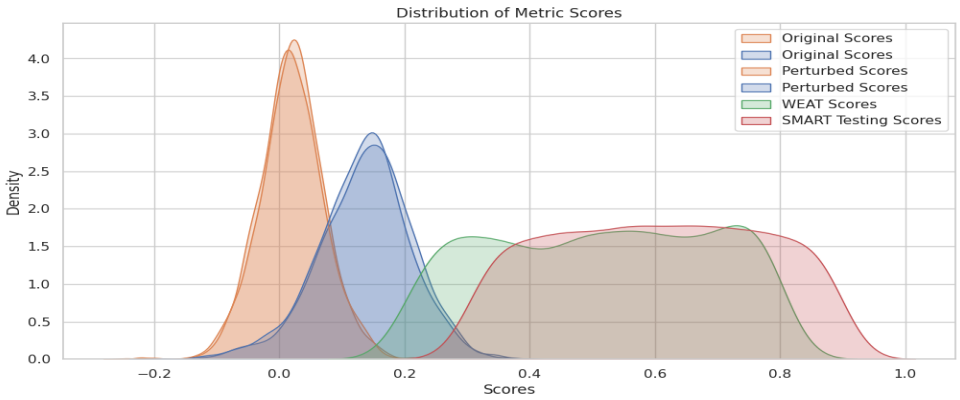

Fig. 2. Density versus scores of various metrics.

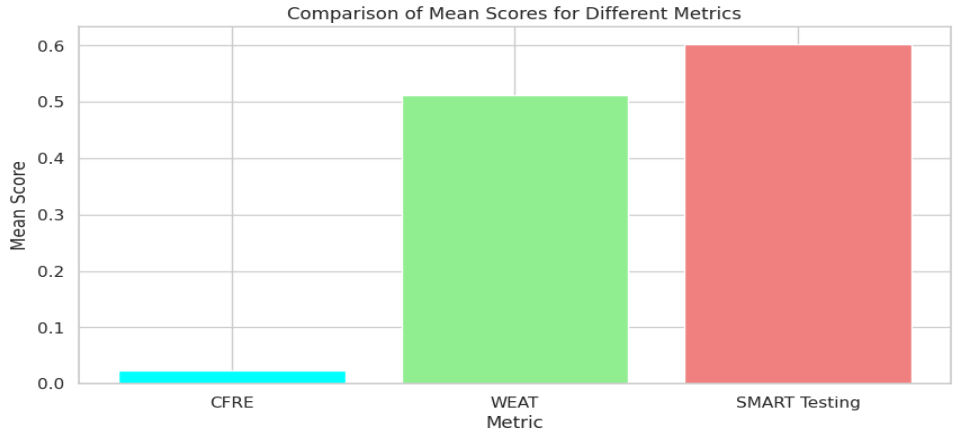

Fig. 3. Mean scores versus various metrics.

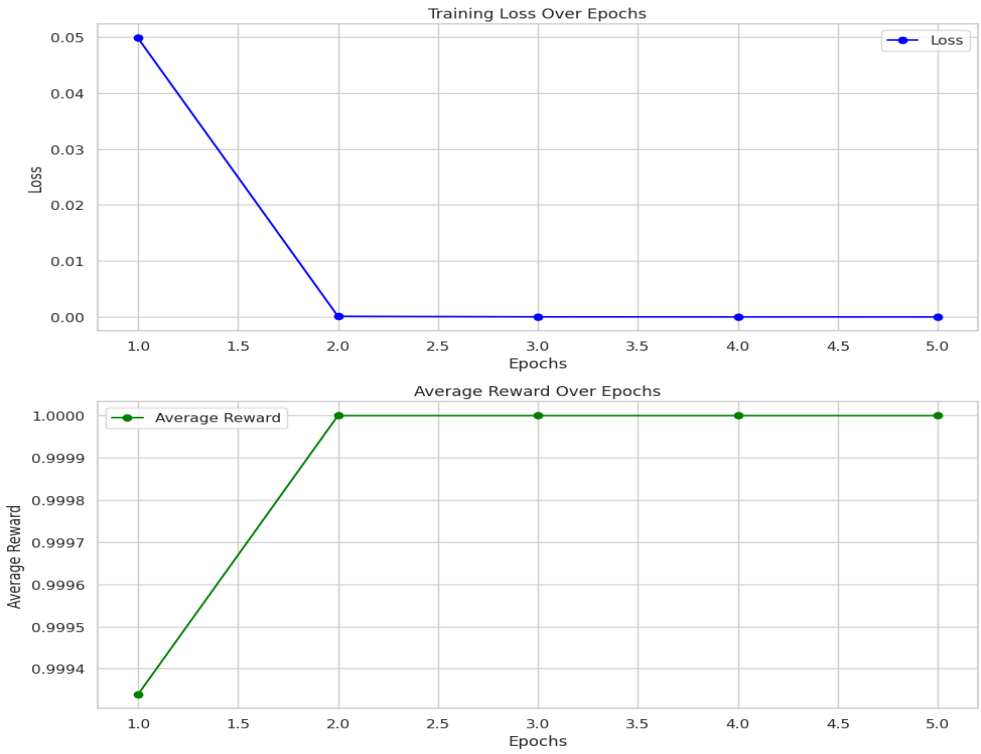

Fig. 4. Graph for loss versus Epochs and Average reward versus Epochs.

VI. CONCLUSION

The integration of the CFRE metric with Self-Imitation Learning (SIL) presents a powerful paradigm for achieving fairness, robustness, and efficiency in reinforcement learning-based NLP systems. This approach ensures that models not only optimize rewards but also address systemic biases, promoting equitable outcomes. By leveraging past successes with fairness-aware constraints, it balances performance and inclusivity, making it especially viable for resource-constrained and real-world applications.

The proposed metric outperformed other existing metrics like WEAT and SMART testing. Also, it got low mean score compared to that of these metrics. The variation between original and perturbed scores serves as a measure of the model's robustness. A narrow difference signifies that the model is resistant to input alterations, showcasing its stability, whereas a wider discrepancy indicates that the model is more vulnerable to adversarial changes or biased modifications in the input data.

REFERENCES

[1] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. https://aclanthology.org/2020.acl-main.442

[2] Bansal, R. (2022). A Survey on Bias and Fairness in Natural Language Processing. *ArXiv, abs/2204.09591*.

[3] Rauba, Paulius & Seedat, Nabeel & Luyten, Max & Schaar, Mihaela. (2024). Context-Aware Testing: A New Paradigm for Model Testing with Large Language Models. 10.48550/arXiv.2410.24005.

[4] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM Conference on Health, Inference, and Learning, pages 151‑159, 2020.

[5] Harini Suresh, Jen J Gong, and John V Guttag. Learning tasks for multitask learning: Heterogenous patient populations in the ICU. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 802‑810, 2018.

[6] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In International Conference on Learning Representations, 2020.

[7] Angel Alexander Cabrera, Minsuk Kahng, Fred Hohman, Jamie Morgenstern, and Duen Horng Chau. Discovery of intersectional bias in machine learning using automatic subgroup generation.In ICLR Debugging Machine Learning Models Workshop, 2019.

[8] Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. Advances in Neural Information Processing Systems, 36, 2024.

[9] Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Navigating data-centric artificial intelligence with DC-Check: Advances, challenges, and opportunities. IEEE Transactions on Artificial Intelligence, 2023.

[10] Oleg S Pianykh, Georg Langs, Marc Dewey, Dieter R Enzmann, Christian J Herold, Stefan O Schoenberg, and James A Brink. Continuous learning AI in radiology: Implementation principles and early applications. Radiology, 297(1):6‑14, 2020.

[11] 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 A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning, pages 5637‑5664. PMLR, 2021.

[12] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. Investigating statistical machine learning as a tool for software development. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 667‑676, 2008.

[13] Lea Goetz, Nabeel Seedat, Robert Vandersluis, and Mihaela van der Schaar. Generalization−a key challenge for responsible ai in patient-facing clinical applications. npj Digital Medicine, 7 (1):126, 2024.

[14] Maire A Duggan, William F Anderson, Sean Altekruse, Lynne Penberthy, and Mark E Sherman.The surveillance, epidemiology and end results (SEER) program and pathology: towards strengthening the critical relationship. The American Journal of Surgical Pathology, 40(12):e94, 2016.

[15] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang.Slice finder: Automated data slicing for model validation. In 2019 IEEE 35th International Conference on Data Engineering (ICDE), pages 1550‑1553. IEEE, 2019.

[16] Svetlana Sagadeeva and Matthias Boehm. Sliceline: Fast, linear-algebra-based slice finding for ml model debugging. In Proceedings of the 2021 International Conference on Management of Data, pages 2290‑2299, 2021.

[17] Adebayo Oshingbesan, Winslow Georgos Omondi, Girmaw Abebe Tadesse, Celia Cintas, and Skyler Speakman. Beyond protected attributes: Disciplined detection of systematic deviations in data. In Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.

[18] Shi, Zijing & Xu, Yunqiu & Fang, Meng & Chen, Ling. (2023). Self-imitation Learning for Action Generation in Text-based Games. 703-726. 10.18653/v1/2023.eacl-main.50.