Bias-Aware Paraphrasing using LSTM and Transformer Models

Medha Srivastava
medhasr@bu.edu

Introduction

Paraphrasing tools are a common, essential assistive tool in modern writing, mainly focused on enhancing the grammar and readability of text while maintaining semantic consistency. However, cognitive biases are inevitable in open writing—even within academic or journalistic contexts where the goal is objectivity. This project explores bias-aware paraphrasing: rewriting input sentences to retain meaning while neutralizing biased language. The main approach of this study is to assess the value of fine-tuning paraphrase models, an LSTM and a transformer, with bias-focused data. The greater goal, possibly beyond this study, would be to use any successful model from this study to create a tool that can provide bias-aware paraphrased sentence suggestions for each sentence while writing.

Method

We trained and compared 3 models:

- An LSTM model pretrained only on paraphrase dataset.
- Same LSTM model fine-tuned with bias-neutralizing dataset.
- T5 Transformer fine-tuned with on same bias-neutralizing dataset.

There were 2 baseline models: a simple LSTM and a GPT4-generated output.

Each model was evaluated on semantic preservation (BLEU, ROUGE) and bias reduction (toxicity change, sentiment change). We used parallel sentence pairs from BEADs and paraphrase datasets to form a combined dataset.

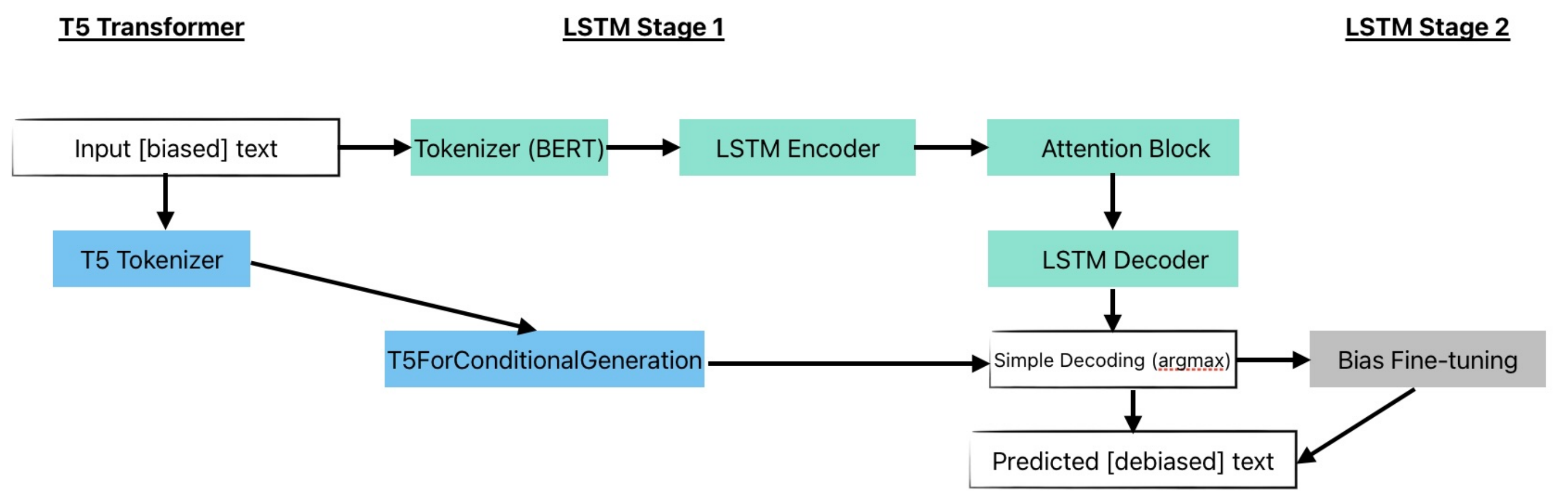


Figure 1: Model architecture, [biased] indicates some inputs with bias focus and others without. Colours represent model-specific blocks, white cells indicate steps shared by all models.

Experiments & Results

The results showed stronger paraphrasing performance in LSTM pretrained-only model, and the T5 model. The fine-tuned LSTM lowered some paraphrase performance metrics and did not improve bias—likely due to an imbalanced dataset with smaller bias samples. The T5 model had higher performance in paraphrase and bias metrics, indicating a strong pretrained model could benefit from fine-tuning given large enough additional data. The GPT baseline also performed relatively well with bias reduction metrics.

The LSTM model showed steady improvements during pretraining (loss dropped from 4.13 to 1.78 over 10 epochs). However, after finetuning on bias data, performance declined in nearly all metrics—suggesting possible overfitting to high-frequency words or the need for more balanced datasets.

The T5 model, by contrast, significantly outperformed all others in both fluency and bias reduction. It generated outputs with higher BLEU/ROUGE and lower sentiment/toxicity shifts, as shown in the side-by-side metric plots in Figure 3.

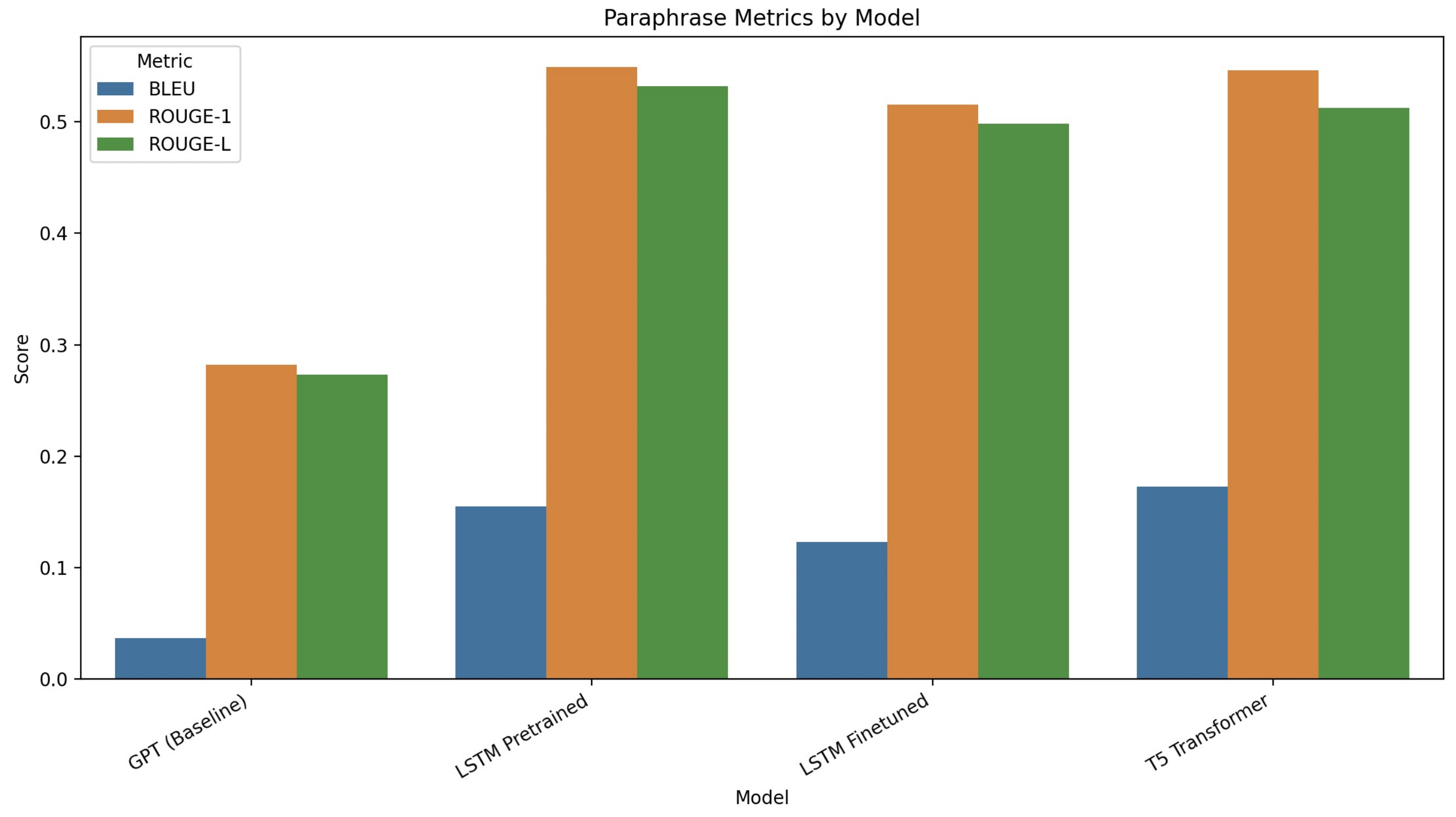


Figure 2: Paraphrase metric results per model.

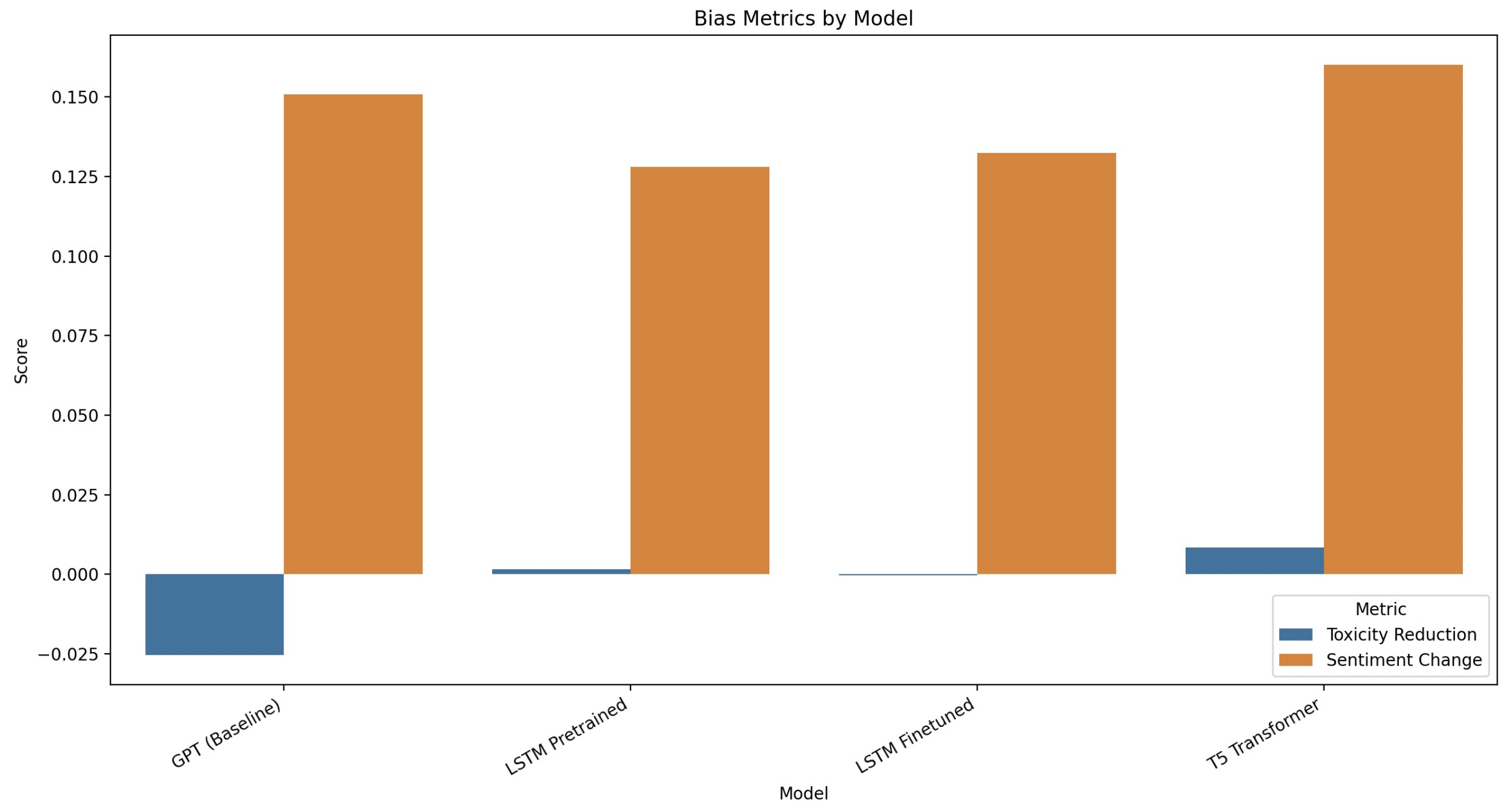
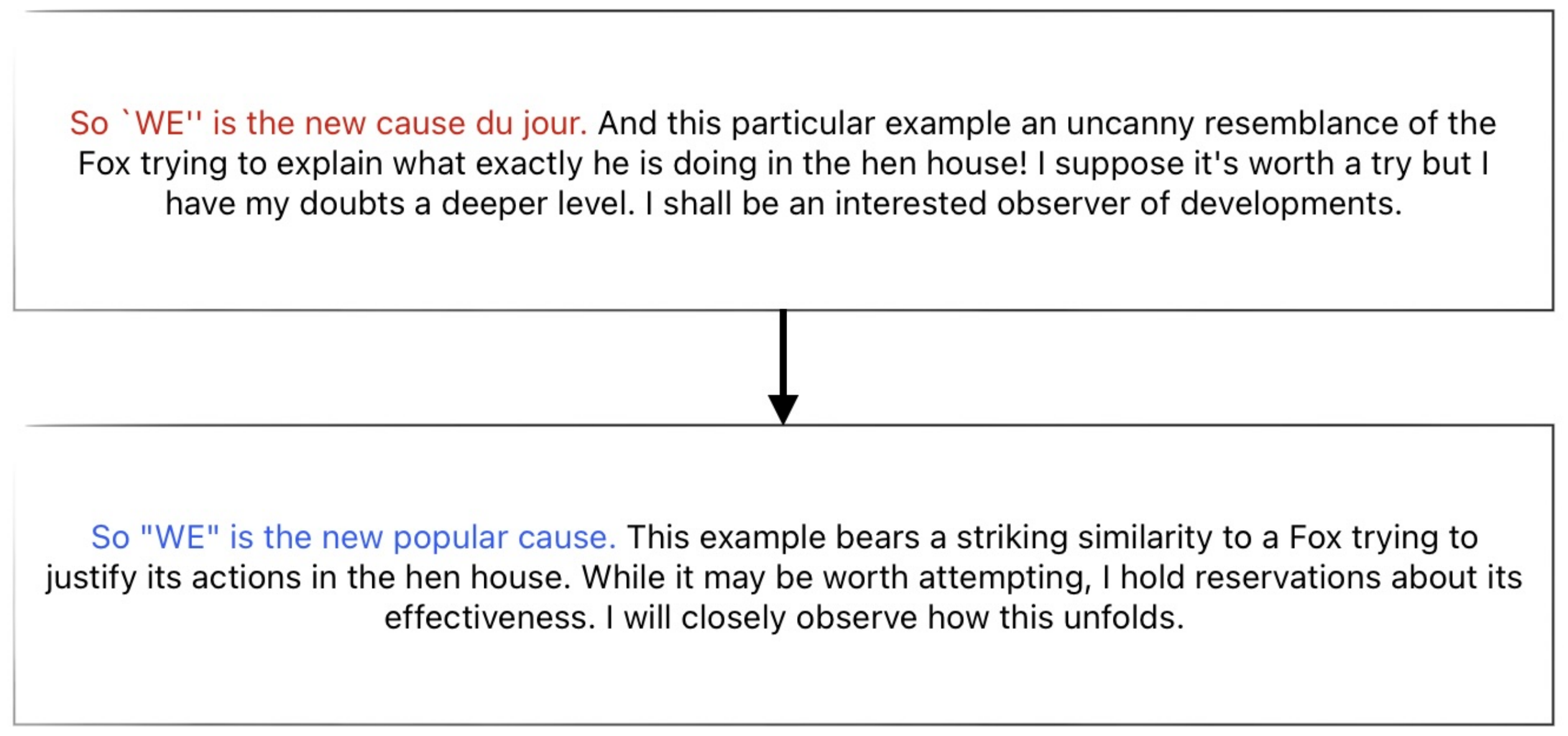


Figure 3: Bias metric results per model.

Conclusion



So `WE'' is the new cause du jour. And this particular example an uncanny resemblance of the Fox trying to explain what exactly he is doing in the hen house! I suppose it's worth a try but I have my doubts a deeper level. I shall be an interested observer of developments.

So "WE" is the new popular cause. This example bears a striking similarity to a Fox trying to justify its actions in the hen house. While it may be worth attempting, I hold reservations about its effectiveness. I will closely observe how this unfolds.

Figure 4: Sample Output from best model (T5) highlighting aim of tool.

Overall, the process of taking a strong pretrained model and fine-tuning it for domain-specific tasks is a valuable approach to integrate more controlled and thoughtful outputs from language generation models. However, fine-tuning with small bias datasets can potentially lower performance—especially if pretrained model is not strong enough independently as our LSTM model showed. In contrast, a pretrained T5 model adapted better to the task with minimal additional training and the small dataset did not lower performance. Results show that strong pretrained transformers can generate fluent, consistent, and more neutral paraphrases—but large, domain-specific bias datasets are crucial for significant improvement. This study is an early step toward building bias-aware rewriting tools.