Reproduction and Extension of "Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation"

Erica Eaton
University of Washington
Computer Science and Engineering

Pirouz Naghavi
University of Florida
Computer and Information Science and Engineering

Introduction

- **Background:** Social bias in natural language data
- **Original work:** “Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation” by Dinan et al., published at EMNLP 2020
- **Model:** ParlAI transformer pre-trained on Reddit conversations
- **Dataset:** LIGHT dialogues
  - Interactions between characters in LIGHT
- **Goal:** Reproducing gender bias mitigation techniques to fine-tune language models
  - Counterfactual data augmentation
  - Positively biased data collection
  - Bias controlled training

Scope of Reproducibility

Evaluation of the following hypotheses made in the original work:

- Combining all 3 bias mitigation techniques yields generated dialogue where percent gendered words and male bias closely match ground truth
- Bias controlled training for the LIGHT dataset yields generated dialogue where percent gendered words and male bias closely match ground truth

Model and Dataset

Model: ParlAI transformer pre-trained on Reddit conversations

- 8 encoder/decoder layers, embedding dimension of 512, and 16 attention heads
- Pre-trained on Reddit conversations, about 2.2 billion samples

Dataset: LIGHT dialogues

- Interactions between characters in LIGHT
- 11,000 interactions and 111,000 utterances
- Dataset variations:
  - Original LIGHT dialogue
  - Counterfactual data augmentation
  - Positively biased data
  - Bias controlled training
  - All bias mitigation techniques are combined

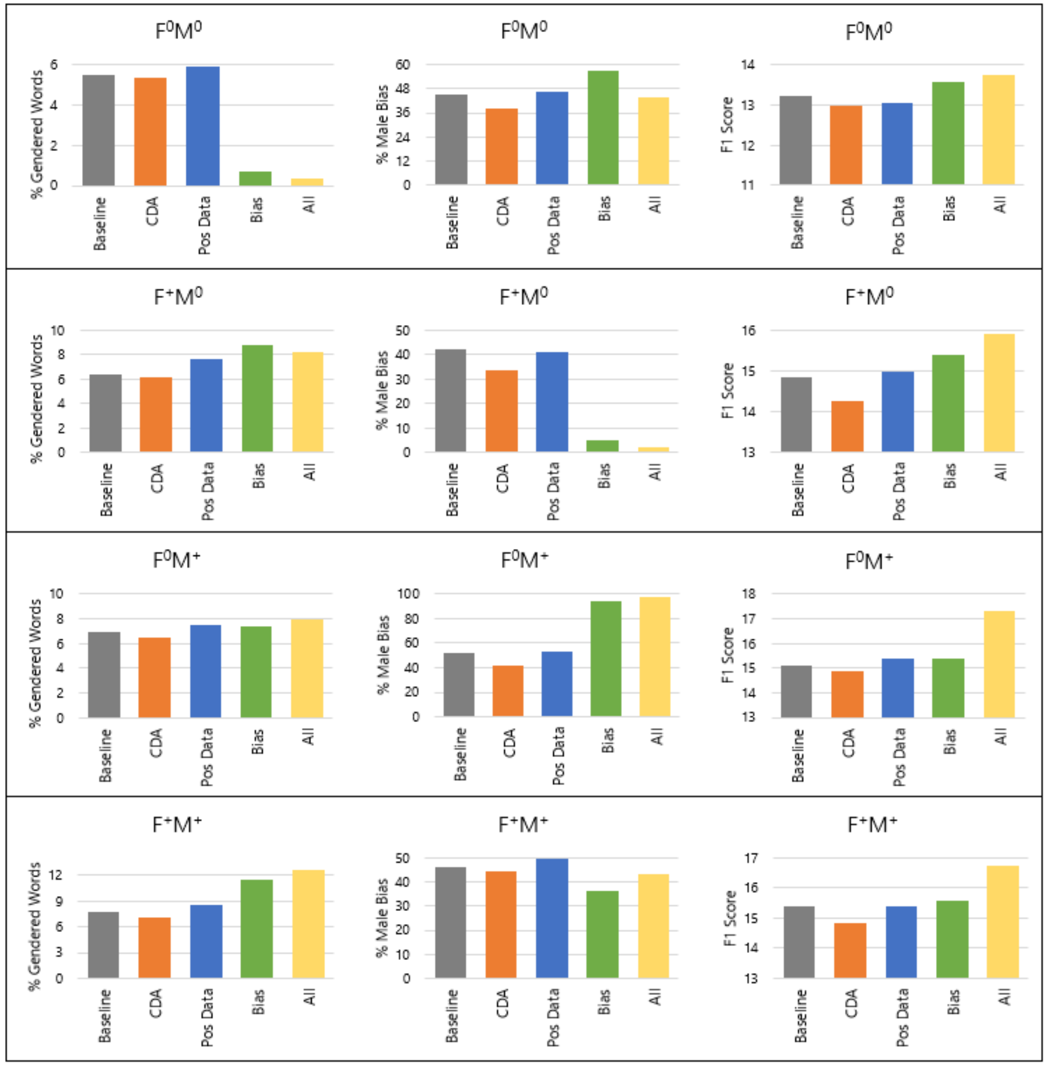
Bias Mitigation Techniques

- **Counterfactual data augmentation**
  - Replace gendered words with their opposite
  - Example: "He is a blacksmith." → "She is a blacksmith."
  - List of 421 gendered words and their opposite
- **Positively biased data collection**
  - Crowd-sourced female character personas and dialogues
  - 507 interactions and 6,658 utterances
- **Bias controlled training**
  - Place dialogues in groups based on number of gendered words
  - Groups: "f0 m0", "f0 m+", "f+ m0", "f+ m+"
  - Group indicator included with dialogue

Reproducibility Results

Results Summary:

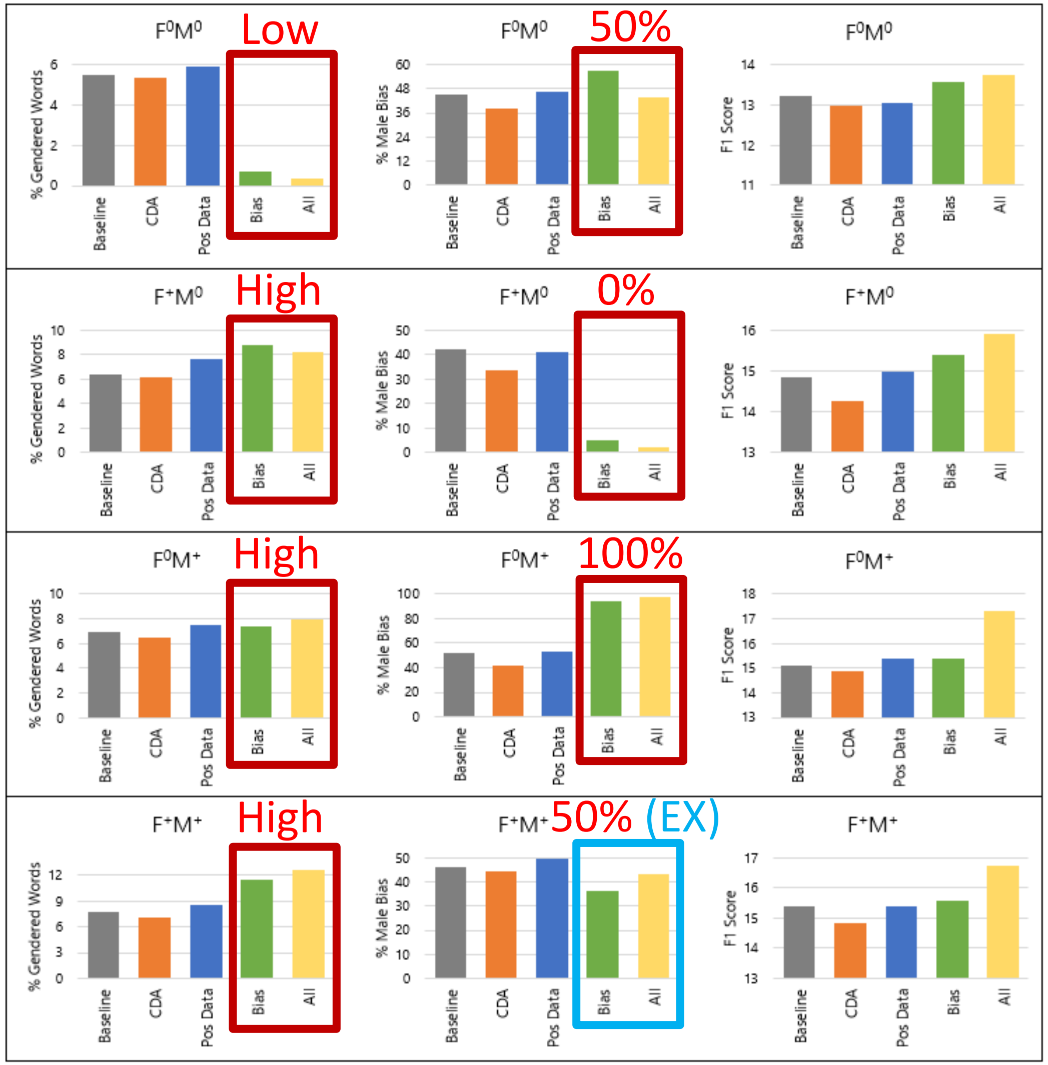
- Results support the reproducibility hypotheses
  - Percent gendered words and male bias similar for “All”, “Bias”, and test labels
  - Exception: male bias for the baseline (46%) is closer to 50% than “All” (43%) and “Bias” (36%) in f+ m+ bin
- Results are quite similar to original work
  - Slight differences in values
  - Main trends are the same
- Main differences between our results and original work:
  - Lower male bias in each bin for the baseline
  - Percent gendered words for “CDA” is closer to the baseline



Reproducibility Results

Results Summary:

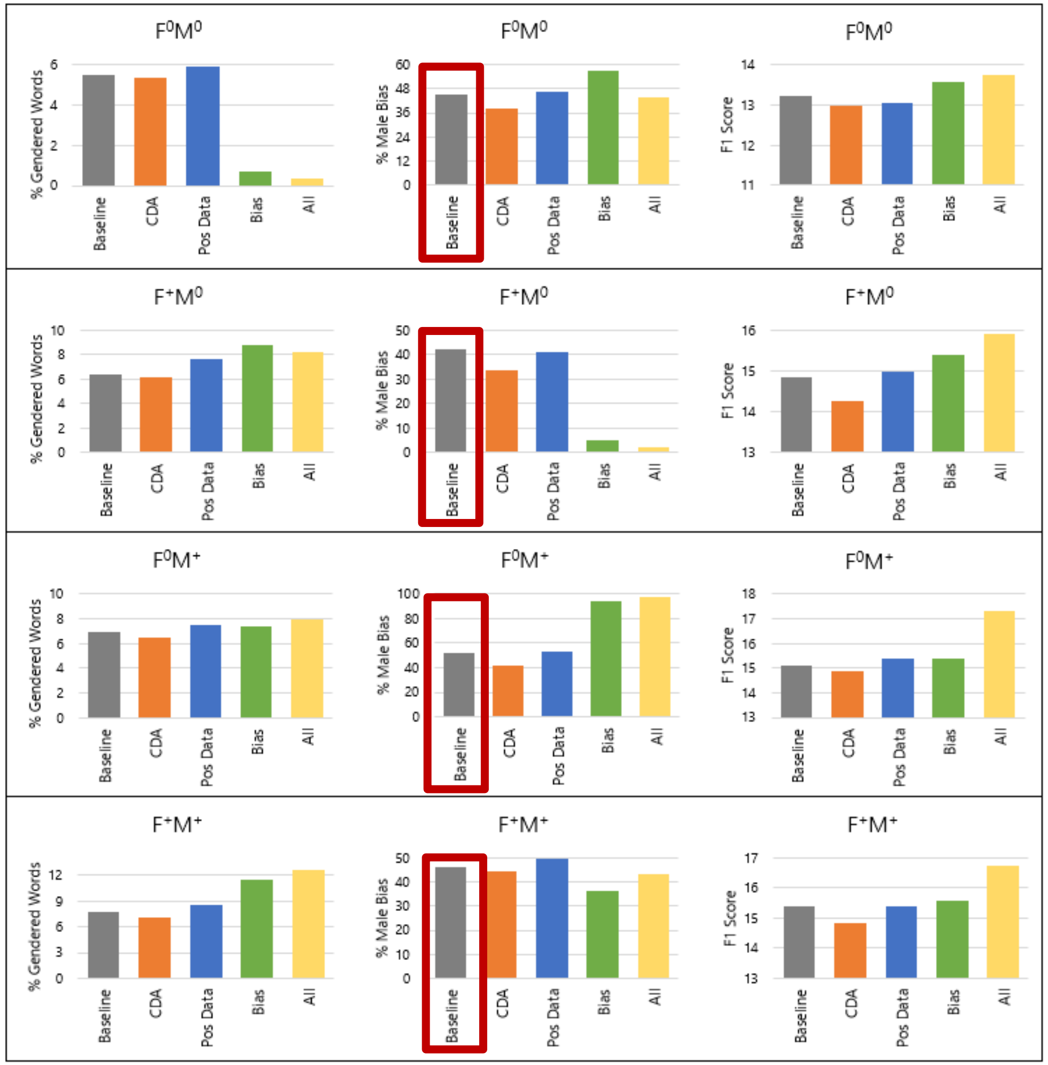
- Results support the reproducibility hypotheses
  - Percent gendered words and male bias similar for “All”, “Bias”, and test labels
  - Exception: male bias for the baseline (46%) is closer to 50% than “All” (43%) and “Bias” (36%) in f+ m+ bin
- Results are quite similar to original work
  - Slight differences in values
  - Main trends are the same
- Main differences between our results and original work:
  - Lower male bias in each bin for the baseline
  - Percent gendered words for “CDA” is closer to the baseline



Reproducibility Results

Results Summary:

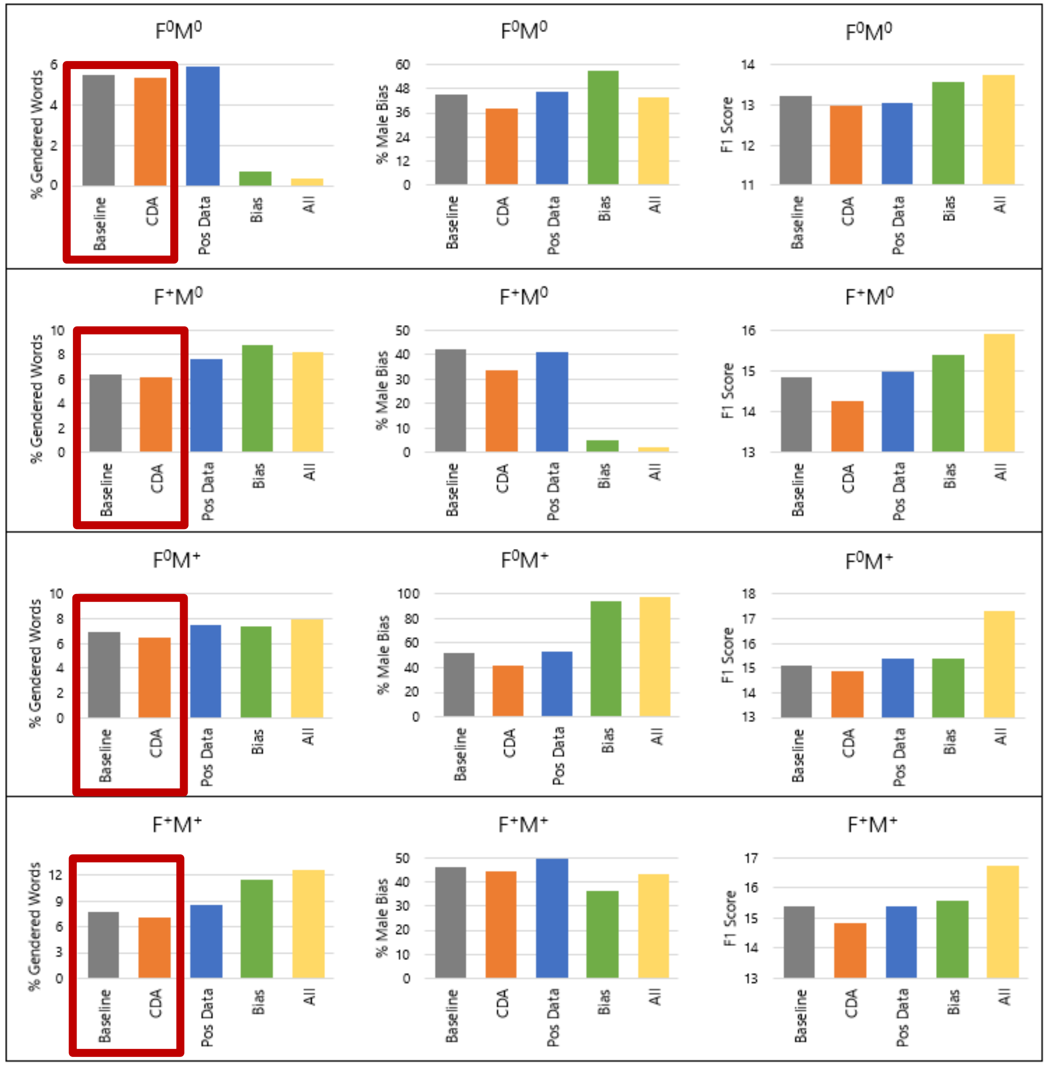
- Results support the reproducibility hypotheses
  - Percent gendered words and male bias similar for “All”, “Bias”, and test labels
  - Exception: male bias for the baseline (46%) is closer to 50% than “All” (43%) and “Bias” (36%) in f+ m+ bin
- Results are quite similar to original work
  - Slight differences in values
  - Main trends are the same
- Main differences between our results and original work:
  - Lower male bias in each bin for the baseline
  - Percent gendered words for “CDA” is closer to the baseline



Reproducibility Results

Results Summary:

- Results support the reproducibility hypotheses
  - Percent gendered words and male bias similar for “All”, “Bias”, and test labels
  - Exception: male bias for the baseline (46%) is closer to 50% than “All” (43%) and “Bias” (36%) in f+ m+ bin
- Results are quite similar to original work
  - Slight differences in values
  - Main trends are the same
- Main differences between our results and original work:
  - Lower male bias in each bin for the baseline
  - Percent gendered words for “CDA” is closer to the baseline



Reproducibility Challenges and Recommendations

1. Pre-trained model
2. Hyperparameters
3. Stopping condition
   - Stopped training when perplexity stopped improving
   - Higher F1 scores than in the original work
   - Potential source for slight variations in results
4. Providing implementation details
   - Paper
   - Website
   - Code
   - Container (resolve dependency issues)

Extensions

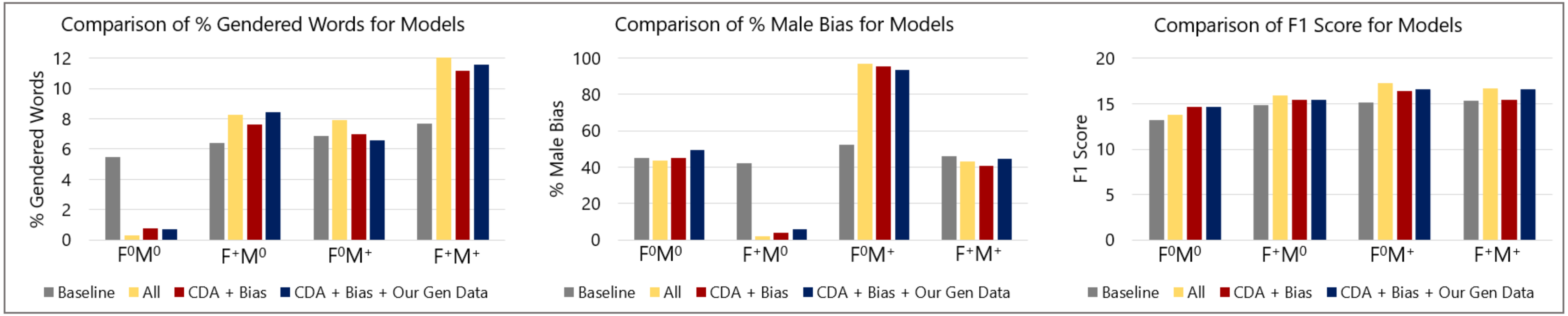
Effect of Removing Positively Biased Data Collection:

- Cost of crowdsourcing data
- Performance loss of excluding the positively biased data

Generating Gender Neutral Data:

- Cheaper alternative to the positively biased data
- Used counterfactual data augmentation and bias controlled training
- Generated responses for all dialogue episodes in the training data
- Neutral model-generated responses used 90% of the time, otherwise actual label is used
- Reconstructed conversations to create a neutral generated dataset
- Fine-tuned the model on new dataset, tested on original test dataset

Extension Results



- "All" achieves better results than "CDA + Bias"
  - Higher F1 scores
  - Percent gendered words and male bias closer to ground truth
- Results for our new model, "CDA + Bias + Our Gen Data," are within 2% of the results for "All"
  - Exception: male bias is closer to 50% than "All" for 3 of 4 bins
- "CDA + Bias + Our Gen Data" yields more gender neutral responses overall

Conclusion

Reproducibility:

- Helpful to provide implementation details in website or paper
  - Model
  - Hyperparameters
  - Stopping condition for training
  - Code or container

Extensions:

- Alternative to crowdsourcing data to make dataset more gender neutral
- Generate dialogue with desired bias using bias controlled training
- Fine-tuned the model on a more gender neutral dataset to help shift generated responses to desired bias

Questions?

Please contact us with any questions.

Erica Eaton
eatone3@uw.edu

UNIVERSITY of WASHINGTON

Pirouz Naghavi
pnaghavi@ufl.edu

UF UNIVERSITY of FLORIDA

FCS Research

Paper link: http://rescience.github.io/bibliography/Eaton_2022.html