Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance

The 2023 Conference on Empirical Methods in Natural Language Processing

Thiemo Wambsganss[1]*, Xiaotian Su[2]*, Vinitra Swamy[2]
Seyed Parsa Neshaei[2], Roman Rietsche[1] and Tanja Käser[2]

[1]Bern University of Applied Sciences, CH
[2]EPFL, Lausanne, CH

Abstract

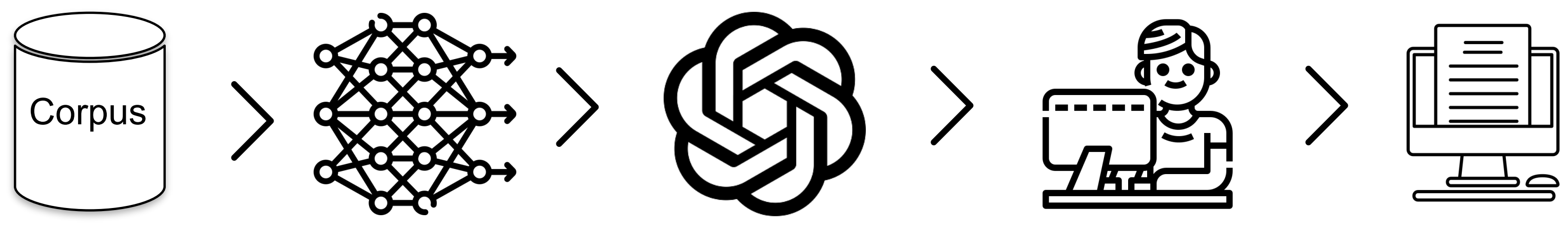
Large Language Models (LLMs) are increasingly utilized in educational tasks such as providing writing suggestions to students. Despite their potential, LLMs are known to harbor inherent biases which may negatively impact learners. Previous studies have investigated bias in models and data representations separately, neglecting the potential impact of LLM bias on human writing.

In this paper, we investigate how bias transfers through an AI writing support pipeline. We conduct a large-scale user study with 231 students writing business case peer reviews in German. Students are divided into five groups with different levels of writing support: **one classroom group** with feature-based suggestions and **four groups recruited from Prolific** -- a control group with no assistance, two groups with suggestions from fine-tuned GPT-2 and GPT-3 models, and one group with suggestions from pre-trained GPT-3.5. Using **GenBit** gender bias analysis, Word Embedding Association Tests (**WEAT**), and Sentence Embedding Association Test (**SEAT**) we evaluate the gender bias at various stages of the pipeline: in model embeddings, in suggestions generated by the models, and in reviews written by students. Our results demonstrate that there is ***no significant difference in gender bias*** between the resulting peer reviews of groups with and without LLM suggestions. Our research is therefore optimistic about the use of AI writing support in the classroom, showcasing a context where bias in LLMs does not transfer to students' responses

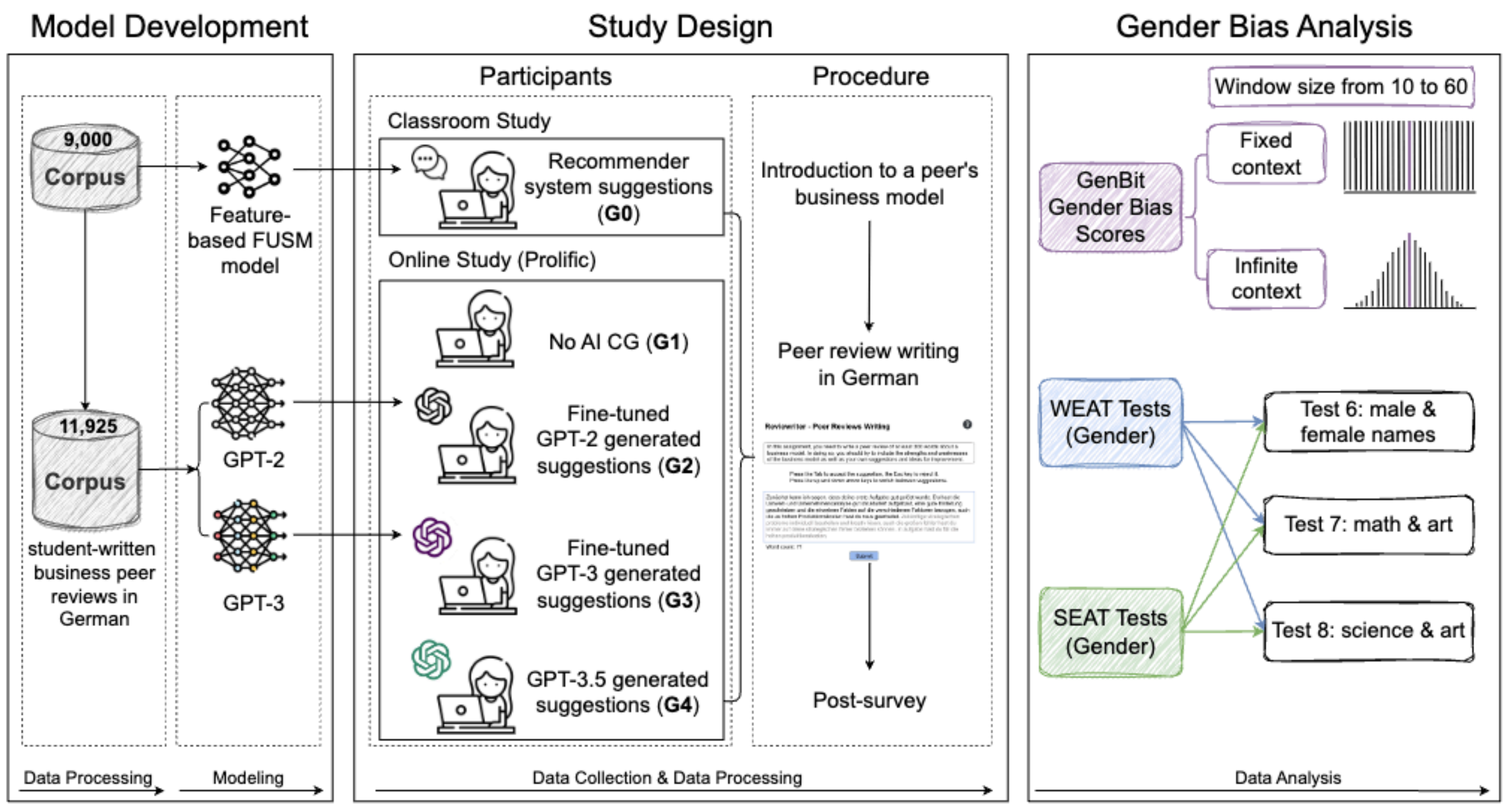
Problem & Solution

? In a real-world peer review writing exercise with AI writing support, does LLM bias transfer to student writing?

? How does bias transfer across the different stages (i.e., model embeddings, model suggestions, student output) of the AI writing support pipeline?



Overview


Model Development
9,000
Corpus
Feature-based FUSM model
11,925
Corpus
student-written business peer reviews in German
GPT-2
GPT-3
Data Processing
Modeling
Study Design
Participants
Procedure
Classroom Study
Recommender system suggestions (G0)
Online Study (Prolific)
No AI CG (G1)
Fine-tuned GPT-2 generated suggestions (G2)
Fine-tuned GPT-3 generated suggestions (G3)
GPT-3.5 generated suggestions (G4)
Introduction to a peer's business model
Peer review writing in German
Post-survey
Data Collection & Data Processing
Gender Bias Analysis
Window size from 10 to 60
GenBit Gender Bias Scores
Fixed context
Infinite context
WEAT Tests (Gender)
SEAT Tests (Gender)
Test 6: male & female names
Test 7: math & art
Test 8: science & art
Data Analysis

Methodology

Data 11,925 student-written business peer reviews in German.

Model Feature Utility Saturation Model (FUSM); Fine-tuned GPT-2; Fine-tuned GPT-3; GPT-3.5

Participants 231 students, controlling for sensitive variables for education level, language, age, and gender

Procedure Watch video on business model -> write reviews with the same tool supported by different models.

Gender Bias Analysis

GenBit

$$P(w|g) = \frac{\text{count}_k(w,g)}{\sum_i \text{count}_k(w_i,g)} \qquad bias(w) = \log\left(\frac{P(w|m)}{P(w|f)}\right)$$

Measuring the association between predefined gendered words and other words via co-occurrence statistics

WEAT

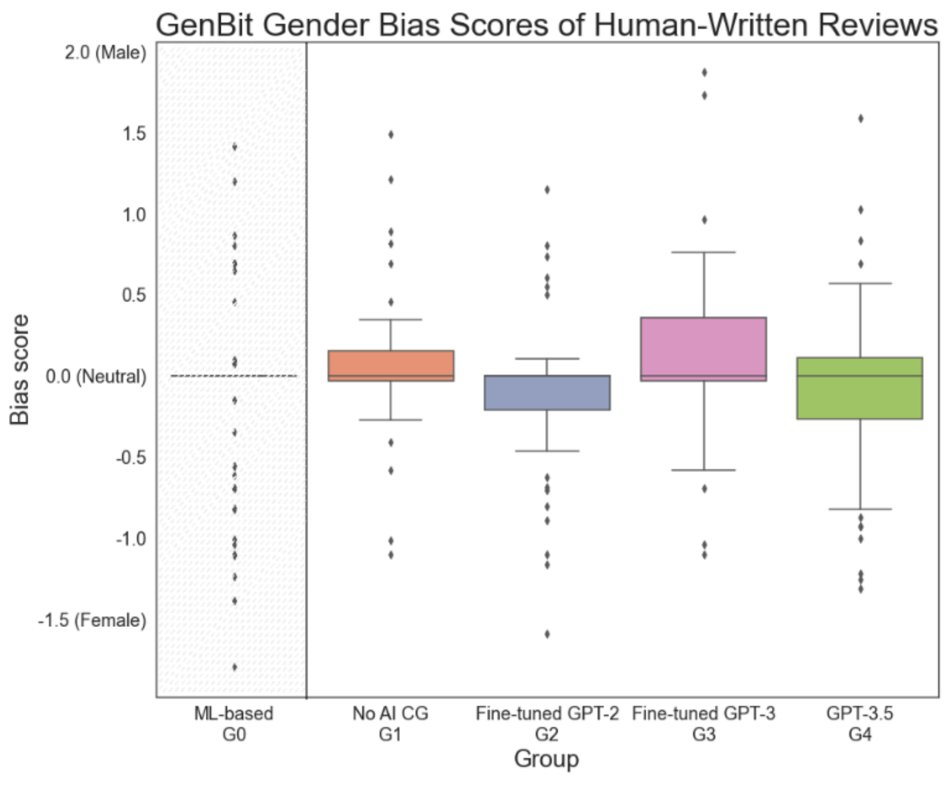
$$\frac{\frac{1}{|X|}\sum_{x\in X} s(x,A,B) - \frac{1}{|Y|}\sum_{y\in Y} s(y,A,B)}{\text{S}_{w\in X\cup Y}(s(w,A,B))}$$

Calculate the semantic similarity between two sets of target words.

SEAT Compares sets of sentences instead of sets of words.

Results

Does bias transfer?

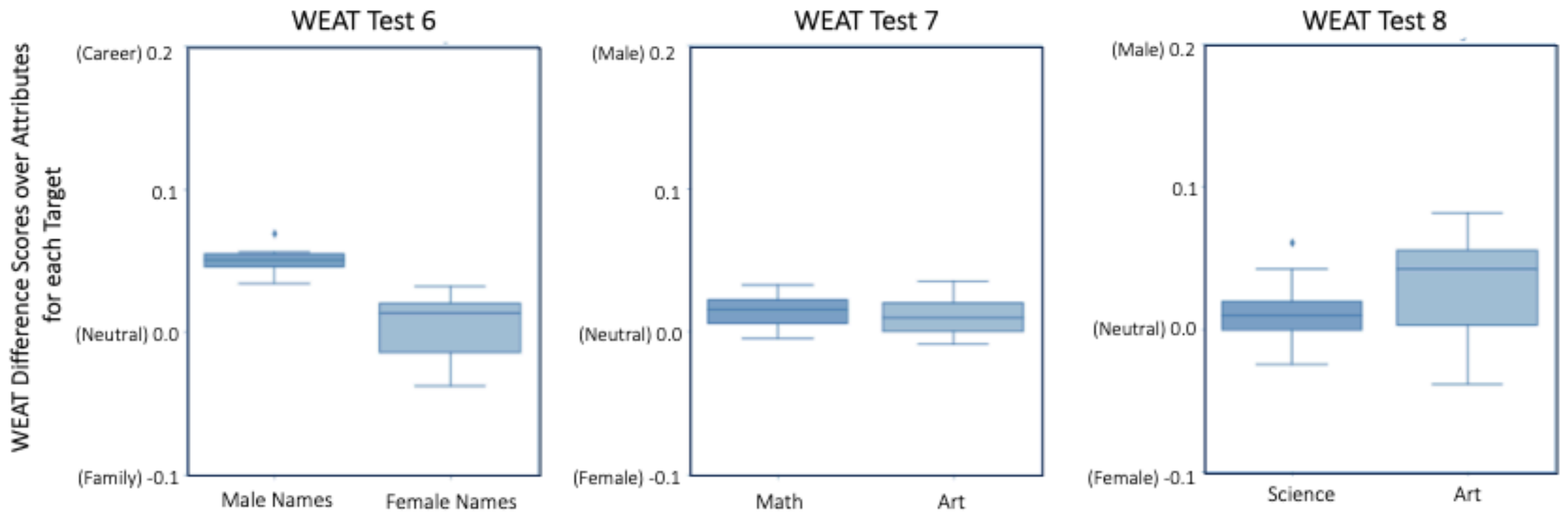


Group	p-value MWU Test GPT-2 (G2)	p-value MWU Test GPT-3 (G3)	p-value MWU Test GPT-3.5 (G4)
Control (G1)	0.170	0.619	0.551
GPT-2 (G2)	-	0.075	0.635
GPT-3 (G3)	-	-	0.269

No statistically significant difference between the bias scores of the four groups.

Students who received writing suggestions from LLMs exhibited the **same** degree of gender bias in their written text as students who received no suggestions.

Where is bias present?



#	Tragets	Attributes	Effect size
6	Male vs. Female Names	Career vs. Family	0.021
6b	Male vs. Female Terms	Career vs. Family	-0.074
7	Math vs. Arts	Male vs. Female Terms	-0.705
7b	Math vs. Arts	Male vs. Female Names	-0.209
8	Science vs. Arts	Male vs. Female Terms	-0.069
8b	Science vs. Arts	Male vs. Female Names	0.078

The gender biases revealed in GPT-2 embeddings did not translate into gender biases in suggestions

Machine Learning for Education Lab
https://www.epfl.ch/labs/ml4ed/

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Parsa Neshaei, Roman Rietsche, Tanja Käser

* Equal contribution

More information contact:
thiemo.wambsganss@epfl.ch
xiaotian.su@epfl.ch

Access our code on GitHub:
https://github.com/epfl-ml4ed/unraveling-llm-bias