Natural Language Processing (NLP) has become increasingly utilized to provide adaptivity in educational applications. However, recent research has highlighted a variety of biases in pre-trained language models. While existing studies investigate bias in different domains, they are limited in addressing fine-grained analysis on educational and multilingual corpora.

In this work, we analyze bias across text and through multiple architectures on a corpus of 9,165 German peer-reviews collected from university students over five years. Notably, our corpus includes labels such as helpfulness, quality, and critical aspect ratings from the peer-review recipient as well as demographic attributes. We conduct a Word Embedding Association Test (WEAT) analysis on (1) our collected corpus in connection with the clustered labels, (2) the most common pre-trained German language models (T5, BERT, and GPT-2) and GloVe embeddings, and (3) the language models after fine-tuning on our collected data-set. In contrast to our initial expectations, we found that our collected corpus does not reveal many biases in the co-occurrence analysis or in the GloVe embeddings. However, the pre-trained German language models find substantial conceptual, racial, and gender bias and have significant changes in bias across conceptual and racial axes during fine-tuning on the peer-review data. With our research, we aim to contribute to the fourth UN sustainability goal (quality education) with a novel dataset, an understanding of biases in natural language education data, and the potential harms of not counteracting biases in language models for educational tasks.

### Methodology

#### Overview of data collection from 2015 to 2019

- **Data collection of peer-review corpus**
  - Submission of Business Models
  - Feedback: Students A, B, C

Students elaborate on the strengths and weaknesses of business models and give recommendations.

#### Word Embedding Association Test (WEAT)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>High (≥ 0)</th>
<th>Low (&lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constructs</td>
<td>5856</td>
<td>3929</td>
</tr>
<tr>
<td>Critical Aspects</td>
<td>5514</td>
<td>3651</td>
</tr>
<tr>
<td>Helpfulness</td>
<td>8086</td>
<td>5279</td>
</tr>
<tr>
<td>Bias</td>
<td>5961</td>
<td>3774</td>
</tr>
</tbody>
</table>

Table 1: Overview of our proposed measured bias categories (conceptual, race, and gender) for the WEAT analysis.

WEAT compares the association between two different target word lists (i.e., Math vs. Arts) to attribute word lists (i.e., Male vs. Female terms). * indicates the original WEAT test number (Caliskan et al., 2017).

### Results

**WEAT Test 6: Gender Bias (Career vs. Family)**

- **High Review Ratings (9-7)**: Helpful
- **Low Review Ratings (1-5)**: Unhelpful

We do not find significant results across any of the nine WEAT tests, with only six co-occurrences identified in total across 9,165 peer-reviews.

**Modeling Bias**

Overall, only one test on the gender axis is able to uncover bias using traditional word embeddings (GloVe).

### More Information/Contact

Thiemo Wambsganss1, Vinitra Swamy1, Roman Rietsche2 and Tanja Käser1

1. EPFL, Lausanne, CH
2. University of St. Gallen, Sankt Gallen, CH

Access our code on GitHub: https://github.com/epfl-mll4ed/bias-at-a-second-glance

Bias at a Second Glance: A Deep Dive into Bias for German Educational Peer-Review Data Modeling

THE 29TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS

Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, Tanja Käser