BAFIS: Dataset + Framework to assess occupational Bias and Human Preference in modern Text-to-image Models

Supplementary Materials:

Thomas Klassert, Adrian Ulges, Biying Fu
RheinMain University of Applied Sciences, Wiesbaden, Germany
Email: Biying.Fu@hs-rm.de

The supplementary material for the paper ”BAFIS: Dataset + Framework to assess occupational Bias and Human Preference in modern Text-to-image Models” provides extra context, data, and insights that support our main findings in the submitted main manuscript. It primarily includes additional figures and algorithms that enhance the understanding of our research. The inclusion of these resources aims to improve transparency and encourage further exploration of the discussed topics. Readers are invited to use these supplementary materials to gain deeper insights and promote ongoing discussions in the field.

Our code and dataset are publicly available on Github. You can find the link here: https://github.com/fbiying87/BAFIS-Occupational-Bias-and-Human-Preference-in-T2I-Models.

A. Appendix to Evaluation Section

In the main script, we discussed the gender distribution of the generated images using different text-to-image models categorized into six occupational categories according to MAGBIG setting.

In Fig. 1, We visualized the gender distributions of model outputs numerically for German prompts only. This focus on German is due to the direct comparison with official occupational statistics published by the Federal Employment Agency for 2024.

B. ELO Rating for Bias Estimation from Human Preferences

As detailed in our main manuscript, Elo rating is used when we want to introduce pairwise comparison evaluation. In our context, we used Elo ratings to compare text-to-image models based on human feedbacks with respect to three aspects. These aspects are bias in terms of gender and ethinicity, percieved image quality and the prompt alignment.

In the listing below, you can see our algorithmic implementation of our Elo implementation.

Algorithm 1 Calculation of the Elo ratings

```
Require: $battles$, $k = 4$, $scale = 400$, $base = 10$, $init\_rating = 1000$
Ensure: Elo ratings of the models
Initialize $rating$ with $init\_rating$ for each model
for all $battle$ in $battles$ do
    Extract $model\_a$, $model\_b$, and $vote$ from $battle$
    $r_a \leftarrow rating[model\_a]$
    $r_b \leftarrow rating[model\_b]$
    Calculation of the expected score:
        $e_a = \frac{1}{1 + base^{\frac{r_b - r_a}{scale}}}$
        $e_b = \frac{1}{1 + base^{\frac{r_a - r_b}{scale}}}$
    if $vote = 0$ then
        $s_a \leftarrow 1$                         ▷ Model A wins
    else if $vote = 1$ then
        $s_a \leftarrow 0$                         ▷ Model B wins
    else
        $s_a \leftarrow 0.5$                       ▷ Draw
    end if
    Update the ratings:
        $rating[model\_a] \leftarrow rating[model\_a] + k \cdot (s_a - e_a)$
        $rating[model\_b] \leftarrow rating[model\_b] + k \cdot (1 - s_a - e_b)$
end for
return $rating$
```

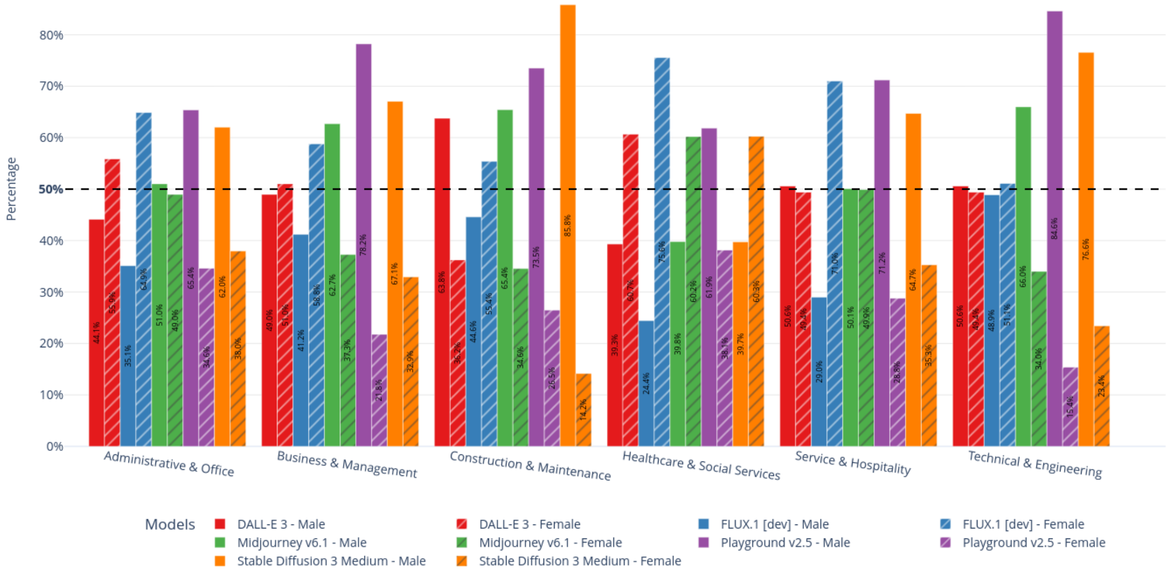

Figure 1. Gender distribution per category and model; German prompts only