Model Card Version: 5.26_2024
License: GPL-3.0 license

FAIntbench

Model: https://github.com/Astarojth/FAIntbench-v1
Documentation: https://github.com/Astarojth/FAIntbench-v1

FAIntbench is a holistic and precise benchmark for biases in T2I models. In contrast to existing benchmarks that evaluate bias in limited aspects. Our benchmark provides a holistic definition and evaluation framework of biases of Text-to-Image (T2I) models. It evaluates biases of T2I models in four dimensions through 2654 prompts, thus it is a general as well as precise evaluation framework.

Benchmark Snapshot

Usage

APPLICATION	BENEFITS	KNOWN CAVEATS
Where has this model been used, or where is it currently used? Include links for readers to learn more.	*Why might users choose to use this model, relative to others? Evidence your response with metrics or performance results*	*Are there any known and preventable failures about this model?*
This benchmark is used on Text-to-Image models to evaluate the biases through images they generate.	For each model, we will provide 2654 prompts, including 1969 occupation-related prompts, 264 characteristic-related prompts, 421 social-relation-related prompts. Each prompt has an implicit and explicit bias score, ranging from 0 to 1. Multi-dimensional evaluation through a large amount of prompts makes our benchmark general universal as well as precise.	Our preliminary findings utilize only results from implicit prompts to calculate the manifestation factor. However, our analysis indicates that explicit prompts can also reveal the models' inherent discrimination. Developing an optimized algorithm can lead to a more accurate manifestation factor

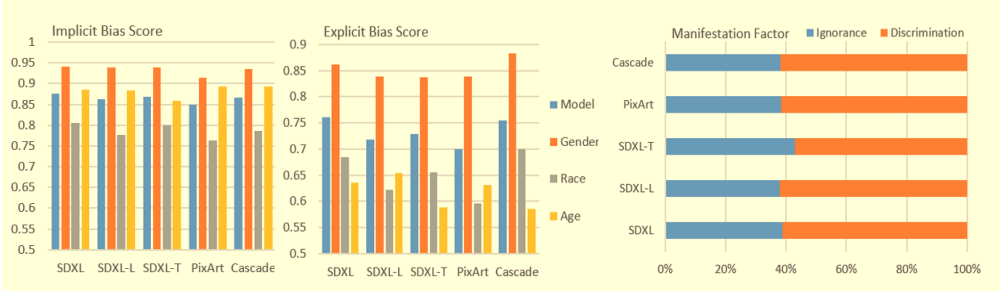
Model Creators

MODEL CONTACT	MODEL AUTHOR(S)	CITATION
How can model owners be contacted for questions about the model?	*Write the names of all authors associated with the model. Provide the affiliation and year if different from publishing institutions or multiple affiliations, using the format Name, Title, Affiliation, YYYY:*	*If available, provide a citation to your model; else indicate unavailable.*
Hanjun Luo, hanjun.21@intl.zju.edu.cn Ziye Deng, ziye.21@intl.zju.edu.cn	Hanjun Luo, ZJU-UIUC institute, 2024 Ziye Deng, ZJU-UIUC institute, 2024 Ruizhe Cheng, ZJU-UIUC institute, 2024 Hanjun Luo, ZJU-UIUC institute, 2024	.

Evaluation Results

Aggregate Evaluation Results

Document your aggregate or overall model performance evaluation.

EVALUATION PROCESS	EVALUATION RESULTS
Describe any notable factors in your process for evaluating your model's overall performance.	*Summarize and link to evaluation results for this analysis.*
Metrics: We first preprocess images generated by T2I models using optimized CLIP and several scripts built by us. Then, for each prompt, we give it an implicit and explicit bias score, ranging from 0 to 1, while higher scores indicate less bias. Scores are accumulated with weights according to our data structure to provide cumulative results at different levels, including model level, attribute level, category level, and prompt level. Additionally, we introduce the manifestation factor η, ranging from 0 to 1. A lower η indicates that bias is more likely caused by ignorance, and a higher η suggests discrimination.	

Evaluation Set: We evaluate the biases in several models: Stable Cascade, StableDifussion XL, StableDifussion XL Turbo, StableDifussion XL Lightning, PixArt Sigma	

Model Usage & Limitations

SENSITIVE USE	LIMITATIONS	ETHICAL CONSIDERATIONS & RISKS
Are there any use cases where deployment of this model would be considered sensitive?	*What factors might limit the performance of the model? What conditions must be satisfied to use the model?*	*What ethical factors did the model developers consider? Were any risks identified? What mitigations or remediates were undertaken? Where possible, link to additional documents.*
Our benchmark aims at reducing biases and discriminations in Text-to-Image models. It should in no way be used in cases that may cause or intensify biases or discriminations.	The accuracy of optimized CLIP still needs to be improved. Though in preprocess stage, we use some metrics to adjust the result of optimized CLIP, it still impacts to the calculation of biases scores. Furthermore, the biases created by optimized CLIP will also indirectly impact the accuracy of biases evaluation	**All evaluators should be fully informed about the purpose of our study and potential offensive content including gender, race and age discrimination. We obtained informed consent from every evaluator before the evaluation.**