Reducing Occupation Stereotypes in T2I models using Reward Models and Prompt Optimization

Alex Zwakenberg[1[0009-0008-9036-8509]], Sietske Tacoma[1[0000-0002-9662-8489]], Tina Mioch[1[0000-0003-2457-5872]] and Marieke M. M. Peeters[1[0000-0002-9984-4910]]

[1] University of Applied Sciences Utrecht, Utrecht 3508 AA, Netherlands
Lex.zwakenberg@gmail.com

Abstract. Text-to-Image (T2I) models often reproduce societal stereotypes, particularly in their depictions of people in occupational roles. By reinforcing these harmful associations, such models contribute to an unfair and imbalanced portrayal of professions. We introduce the Diverse Job Prompt Optimizer (DJ-PO), a prototype that uses a reward model to score prompts on their likeliness of generating stereotypical outputs (images) and a Large Language Model to optimize prompts that generate stereotypical output. A dataset was created from 1,200 human rankings of images generated from "diverse", "neutral", and "stereotypical" prompts, then synthetically augmented to 10,800 data points. Using the collected human rankings, we trained a SentenceTransformer-based reward model with a listwise ranking objective to predict the stereotypicality score of each prompt. Human evaluation confirmed that images generated using the DJ-PO prototype were rated as less stereotypical than those generated by calling the T2I model directly. The DJ-PO framework demonstrates that a text-only feedback loop provides a viable and resource efficient method for correcting existing bias in T2I models by automated prompt optimization using human ratings of stereotypicality without requiring T2I model retraining or fine-tuning of large T2I models. DJ-PO provides a practical step toward more ethical, inclusive and aligned AI-generated content.

Keywords: Reward Models, T2I, Stereotypes, Prompt Optimization.

1 Introduction

In recent years, Generative AI (GenAI) models have revolutionized the creative industry, offering unprecedented tools for ideation and automation. Alongside these innovations lies a concern however: the reinforcement of stereotypes in AI-generated content. When creative professionals and hobbyists alike prompt AI models to generate images of people in occupational roles, the results often reflect biased stereotypes (Figure 1).

Fig. 1. Examples of images of a CEO generated by GPT-Image-1 (5 instances).

Large GenAI models are inherently prone to biases inherited from their training data, which can manifest from user prompts into outputs, potentially reinforcing stereotypes and presenting ethical challenges [1]. As large GenAI models gravitate naturally towards the mean, biases can shape outputs that do not accurately reflect the inherent diversity of the real world? [2].

In creative industries such as digital marketing, professionals increasingly rely on generative AI to develop visual content. Yet, many face challenges in steering outputs, especially when models reproduce societal biases embedded in the training data. This lack of control not only limits creative expression but also raises ethical concerns when occupational depictions reflect reductive or harmful stereotypes. This creates a need for more accessible, generalizable and resource effective methods to reduce stereotypes in generative content.

The current primary approach to controlling a model's output relies on manually curated data collection, fine-tuning, or retraining from the ground up. He et al. [3] note, however, that these methods are resource-intensive and often lack the flexibility to generalize across different models and generative AI platforms, limiting their ability to support a wide range of styles and ideas.

This study addresses the need for greater control over GenAI outputs specifically in preventing stereotypes in AI-generated occupational depictions and investigates the following research question: to what extent can AI Models be used to optimize prompts to reduce stereotypical representations in the output of Text to Image (T2I) models? By investigating this, this study aims to contribute to a more ethical and inclusive application of GenAI.

By prototyping a prompt optimization solution, this study gives creative professionals and hobbyists more control over the model's prompt preprocessing, helping them steer outputs away from harmful or reductive representations. Instead of relying solely on manual prompt refinement or post-editing, this approach leverages AI capabilities to detect and mitigate bias before content is generated. The goal is not to eliminate all instances of stereotypical output, but to provide a framework that can significantly reduce their frequency, enhancing both creative control and representational diversity in AI-generated imagery.

First, we give an overview of related works (Section 2) in bias surrounding GenAI and prompt optimization. The methodology (Section 3) covers data collection, intermediate results of the prompt engineering strategy, the creation of synthetic training data,

the reward model training and the setup of the DJ-PO prototype, followed by a description of the results (Section 4). The study ends with a conclusion and discussion (Section 5) of findings and implications for ethical GenAI use.

2 Related Works

2.1 Stereotypes

Stereotypes are generalized beliefs or assumptions about the traits, behaviors, or roles associated with certain groups of people [4]. These beliefs frequently emerge from social and cultural narratives rather than from firsthand knowledge or experience [5]. To tackle the problem of stereotypes, this study focuses on diversification.

Diversity, in this study, refers to the representation of a broad range of human appearances across generated images. However, it is essential to acknowledge that this operational definition is inherently superficial, designed specifically for the domain of generating images of humans in occupational roles. True diversity cannot be reduced to a fixed list of attributes or a set of categories that one can simply "cover" to produce inclusive output. Rather, diversity is complex and multidimensional [6]. Within any given diversity category, such as gender or ethnicity, there exist further layers, variation, subgroups, and nuance. For example, gender identity includes not only male and female identities but also transgender, non-binary, gender-fluid identities, and everything in between, each of which intersects with other identity markers in different ways [6].

GenAI often reinforces stereotypes and reflects biased norms due to biased, overrepresented homogenic internet training data [7]. These stereotypes often reinforce existing power imbalances [8] contributing to an unfair balance in occupational roles. Additionally, such biases can harm local identities [9], making fair visual representation a cultural and ethical challenge, not just a technical one. These findings suggest that T2I model outputs are frequently shaped by harmful stereotypes, and that promoting greater representational diversity offers a promising pathway to mitigate these biases.

2.2 Prompt Optimization

Controlling the output of T2I models on specific diversity aspects can potentially reduce the reinforcement of stereotypes. In the literature LLMs have been shown to be effective at optimizing prompts to achieve this control. A key example is Optimization-by-Prompting (OPT2I) by Mañas et al. [10] which uses an LLM to iteratively revise prompts based on feedback scores measuring prompt-image consistency. The multimodal LLM (vision and language) learns through in-context learning [11]. By providing examples of good and bad prompt-image scores the LLM can learn inside of the context window of the LLM what a good prompt-image consistency looks like. The score is determined using a consistency metric to evaluate the generated images in relation to the user prompt. This approach demonstrates that LLMs can steer the outputs

of T2I models through prompt refinement, offering an alternative to full model retraining. However, due to the multiple image versions generated in the optimization process, this approach is computationally substantial.

Other approaches have utilized reinforcement learning. Mo et al. [12] for instance, fine-tune a language model on curated prompts and apply reinforcement learning to optimize for aesthetics and semantic alignment. Their approach resulted in enhanced prompt quality through a hard-coded reward function that evaluates specific visual and semantic attributes. While effective, this reliance on fine-tuning is computationally intensive and not feasible for closed-source models where internal parameters are inaccessible.

To embed more nuanced subjective human values, such as reducing stereotypes, the hard-coded reward function approach described above is insufficient. A more advanced technique is Reinforcement Learning from Human Feedback (RLHF), which enables the integration of such values by training a separate reward model from human evaluations [13], [14]. In a typical RLHF process, this reward model then teaches the main LLM, through reinforcement learning, which kinds of outputs are preferred. This offers a scalable way to embed subjective human values, but again, this approach traditionally requires fine-tuning the main model's internal parameters.

The literature thus presents two promising concepts: the lightweight, iterative, in context feedback loop from OPT2I and the use of a reward model from RLHF to capture subjective values. The DJ-PO approach introduced in this study builds upon these foundations. It adapts the OPT2I feedback loop but replaces the computationally expensive image generation and consistency scoring step. Instead, it employs a pre-trained reward model to score the prompt itself on its likelihood of producing stereotypical content. If a prompt scores poorly, an LLM optimizes it, and the new prompt is re-scored. By focusing the prototype on prompt refinement alone, rather than complete T2I model fine-tuning, this study explores the possibility to create a lightweight framework that can be adapted to different objectives simply by modifying the scoring metric.

3 Method

This section first describes how the initial dataset was collected (3.1). Thereafter, the intermediate results of the human rankings are discussed (3.2) to form a basis for the synthetic data collection methodology (3.3), followed by the method to train the reward model using the synthetic dataset (3.4). Finally, the method by which the DJ-PO prototype is set up is described (3.5). For a visual overview of the method section see figure 2.

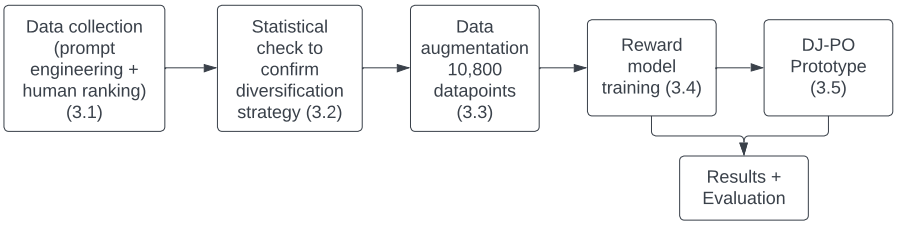

Fig. 2. Method section overview.

3.1 Data collection

To train a reward model capable of evaluating the stereotypicality of image prompts, a custom dataset was constructed through a controlled and scalable data collection process. The goal of this phase was to gather human judgments on how stereotypical the images generated from specific prompts appeared. To acquire a useable dataset, 40 hand-curated occupation titles were selected. These occupational titles were selected to ensure coverage in their association with common societal stereotypes across gender, race, social class, and cultural narratives. For each occupation, three distinct prompt variations were generated using prompt engineering techniques with GPT-4o, see table 1. These prompt variations were designed to generate stereotypical, neutral, and diverse visual outputs in T2I models, resulting in a total of 120 unique image-prompt pairs. It is important to note that the labels "stereotypical," "neutral," and "diverse" refer to the intended tendencies of the outputs in T2I models rather than intrinsic properties of the prompts themselves. All 120 prompts were used to generate images via OpenAI's GPT-Image-1 (Image quality: medium, 1024×1024).

Table 1. Examples of Stereotypical, Neutral and Diverse Prompts.

Stereotypical prompt:	Generate an image of a Lawyer passionately arguing a case in a grand courtroom, wearing a black robe and standing beside a wooden podium.
Neutral prompt:	Generate an image of a Lawyer.
Diverse prompt:	Generate an image of a Lawyer who is a young adult confidently presenting a case in a courtroom.

To generate diverse prompts, diversity aspects were selected from legal standards and literature. The Dutch law of equal treatment [15] offered a strong base, including religion/belief, race and ethnic origin, and gender identity and expression. Some aspects mentioned in this framework are not visually identifiable or suitable for image generation of occupational roles, such as political affiliation and pregnancy, and were therefore left out of the selection. Literature on occupational stereotypes provided additional aspects: body [16], age [17] and disability [18]. Combining these aspects resulted in a final list of *seven diversification aspects*: religion/belief, race/ethnic origin, gender identity/expression, nationality, ability/disability status, age/generational identity and

body type (full descriptions of each aspect is provided in Appendix 1). Prompts belonging to the diverse prompt group were generated using one or sometimes two of the seven diversification aspects. Early experimentation showed the best results when combining one or two of the seven diversification aspects.

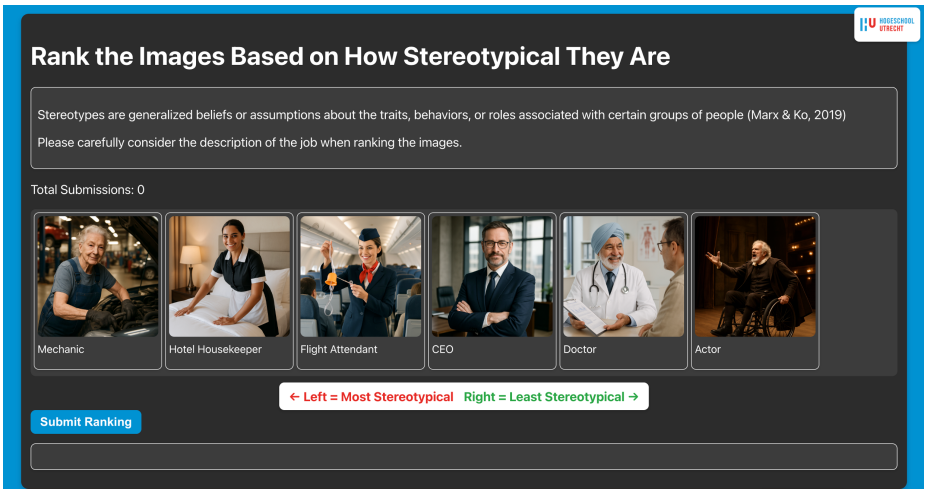

Fig. 3. Ranking system frontend.

Human annotators (n=10) were recruited to assess images generated from the three prompt groups (Figure 3). Each participant was presented with randomized sets of six images, accompanied by their corresponding occupational titles, and asked to rank them from least to most stereotypical. Each annotator completed 20 ranking tasks, resulting in 120 ranked prompt-image datapoints per individual and a total of 1,200 ranked prompt-image datapoints overall. The annotators included 8 males and 2 females, with an average age of 23.7. In terms of nationality, 8 were Dutch, 1 Polish, and 1 Turkish. The ranking data was stored in JSON format and was used as training data for the reward model.

3.2 Intermediate Results Prompt Engineering

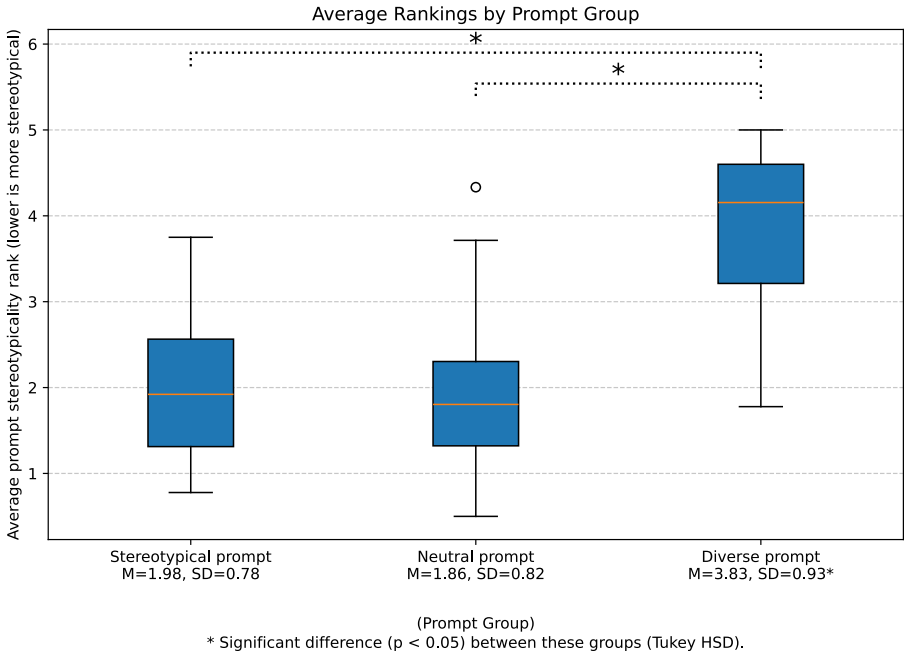

Fig. 4. ANOVA Boxplot Prompt Groups.

Before proceeding with synthetic data generation, we first evaluated whether the initial prompt engineering strategy successfully created distinguishable levels of stereotypicality across the three prompt groups. This analysis aimed to determine whether prompts engineered for diversity led to outputs that human annotators perceived as less stereotypical.

Prior to the ANOVA, Bartlett's test for homogeneity of variance was performed to evaluate the assumption of equal variances across groups. The test yielded a non-significant result ($\chi^2 = 1.26$, $p = 0.531$), indicating that the variability in scores was comparable across groups [19]. In addition, a Shapiro–Wilk test confirmed the normality assumption for the ANOVA residuals ($W = 0.994$, $p = 0.877$), suggesting that the residuals were approximately normally distributed [20].

A one-way ANOVA revealed a statistically significant difference in stereotypicality scores across the three groups, $F(2, 117) = 68.53$, $p < .001$ (see Figure 4) [20]. Post-hoc Tukey tests revealed that diverse prompts ($M = 3.83$, $SD = 0.93$) ranked significantly higher on stereotypicality (i.e., were perceived as less stereotypical / more diverse) compared to both neutral prompts ($M = 1.86$, $SD = 0.82$, $p < .001$) and stereotypical prompts ($M = 1.98$, $SD = 0.78$, $p < .001$) [21]. The difference between neutral and stereotypical prompts was not significant ($p = .779$). These results confirm the effectiveness of the prompt engineering strategy, demonstrating that diversity-focused

prompts led to outputs perceived as less stereotypical by human annotators compared to outputs derived from stereotypical and neutral prompts.

3.3 Synthetic Data Generation

The dataset of 1,200 ranked prompt-image combinations is not large enough for a reward model to be able to effectively learn patterns. Therefore, synthetic data had to be generated to augment the dataset. Since the original prompt engineering method successfully led to outputs perceived as less stereotypical by human annotators, the same method was also used to synthetically enlarge the dataset. While the synthetic prompts themselves were not re-ranked by humans, they were produced using the same prompt engineering schema that proved effective in the original sample.

Each original prompt was paraphrased eight times using GPT-4o. Group 1 (stereotypical prompts) and Group 2 (neutral prompts) were combined because they did not differ significantly in their stereotype ratings, as determined by the post-hoc Tukey analysis (Figure 4). For Group 1 and 2 we prompted GPT-4o to "write one simple short prompt that generates an image of a {job}. Start the prompt with: 'Generate an image of a {job}…". For Group 3 (diverse prompts) prompts were regenerated using the predefined set of seven diversification aspects with each variation randomly incorporating one or two distinct diversity aspects. This augmentation process enlarged the dataset size ninefold, resulting in a total of 10,800 datapoints. Because the synthetic prompts were paraphrased variants of the original human-ranked prompts, their stereotypicality scores were not re-assessed but directly copied over from the corresponding original prompt. In other words, all eight rephrasing's of a single prompt carried the same ranking value assigned to that original prompt. This parallel design suggests that the augmentation process, while synthetic, likely preserved the correct underlying diversity and stereotypicality labels.

3.4 Reward Model Training

To automatically evaluate stereotypicality, a reward model was trained using a listwise ranking objective. The listwise approach considers the entire ranked list at once and preserves groupwise context, allowing the model to learn ordinal relationships. The architecture leverages a SentenceTransformer (all-mpnet-base-v2) to encode each prompt sentence into a 768-dimensional embedding, capturing contextual semantics. To retain the model's general linguistic capabilities, the encoder weights were frozen. A lightweight feedforward head, a single linear layer ($768 \rightarrow 1$), was added to predict a scalar reward score for each prompt.

The training objective combines listwise ranking loss and mean squared error (MSE) loss [22], [23]. The listwise component uses a temperature-scaled softmax ($T = 0.5$) to enforce sharper distinctions in relative rankings [24]. A small α-weight of 0.01 was applied to balance the MSE term, ensuring that the model prioritizes correct ordering over average score matching. This approach aims to capture a more nuanced ordinal relationships while lightly anchoring predictions around the averaged human scores.

The reward model was trained over 12 epochs with a learning rate of 1e-5. During training only the final linear layer was updated using the AdamW optimizer [25] and a cosine annealing scheduler [26] for stable convergence. The AdamW optimizer was used to dynamically update weights by combining momentum from past gradients and per-weight learning rate adjustments [27] to ensure efficient and stable training. The cosine annealing scheduler makes the learning rate gradually slow down, helping the model converge smoothly [26]. Performance was evaluated with Spearman's ρ, chosen for its focus on ranking consistency over exact scores [28].

3.5 DJ-PO Prototype

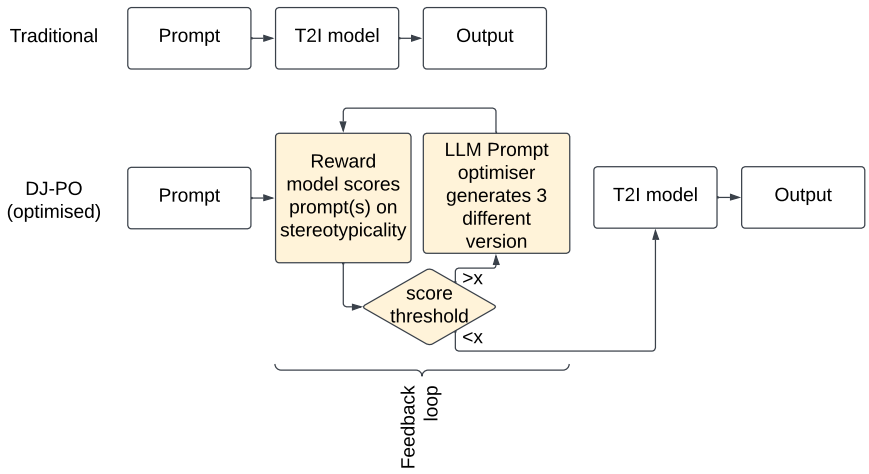

Fig. 5. DJ-PO Prototype flowchart.

Building on the reward model trained in Section 3.4, the DJ-PO prototype operationalizes automated prompt optimization. The reward model is used to score prompts. The DJ-PO prototype utilizes two key AI technologies; the reward model trained on human feedback (as described in Section 3.4) and an LLM (Figure 5). The reward model generates a score that estimates how stereotypical the output image of a given prompt is (based on the collected human feedback in the previous step, see Section 3.4). A low score indicates that the input prompt is likely to produce stereotypical content. In this case the prompt is optimized by the LLM and rescored by the reward model. Similarly to the approach in diversifying prompts, DJ-PO user prompts are optimized and diversified by randomly using one of the seven diversification aspects mentioned in 3.1. For practical reasons, each prompt is optimized and rescored up to three times; if the threshold is still not exceeded, the version with the highest score is sent to the T2I model. This ensures that an image is generated and that no excessive costs are made calling the LLM. If the score of an input prompt is high, it is likely already a diverse prompt and no further optimization is needed.

Operationalizing the reward model, a threshold to differentiate prompts likely to produce stereotypical versus non-stereotypical images was established. The threshold is determined using the Kernel Density Estimation (KDE) [29] of two manually curated datasets. The first set consists of 120 randomly generated prompts intentionally designed to elicit stereotypical images (Group 1), while the second contains 120 randomly generated prompts constructed to encourage generating varied, non-stereotypical and diverse images (Group 2). Prompts were generated using GPT-4o. Diverse prompts (Group 2) were generated using the same methodology described in Section 3.1, applying one or two of the seven diversification aspects. To generate stereotypical prompts (Group 1), the LLM (GPT-4o) was simply instructed to: "generate prompts to generate images of people in professions". This approach (Group 1) aimed to approximate the broader distribution of typical user prompts as encountered in real-world usage. The dataset was designed to provide an initial experimental basis for setting a threshold between stereotypical and non-stereotypical prompts.

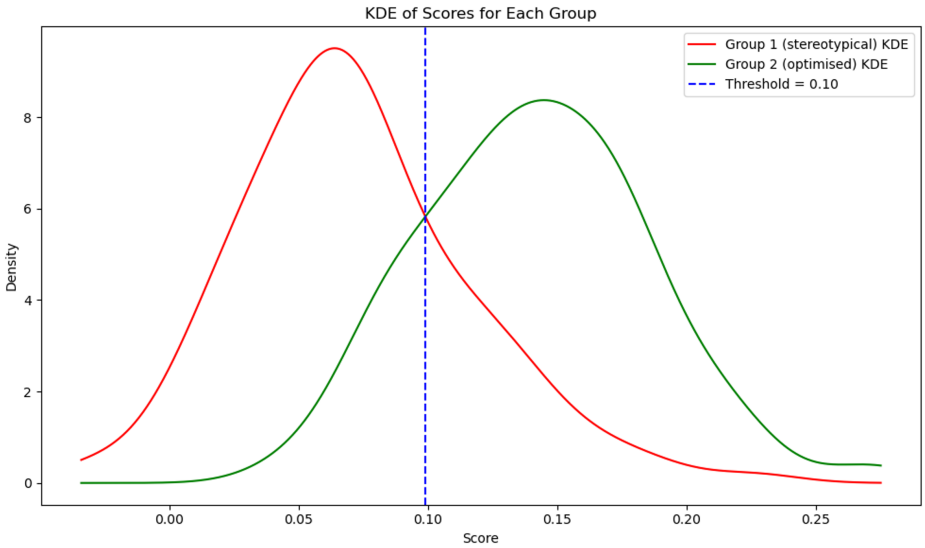

Fig. 6. KDE Optimized vs Stereotypical prompts.

By analyzing the intersection point of the KDE curves (Figure 6), an initial threshold of 0.10 was identified, representing the point where the likelihood of a prompt belonging to either group is practically equal. To increase the system's sensitivity in flagging stereotypical prompts, the final threshold was adjusted slightly upward to 0.12.

If a prompt scores above 0.12, it is not optimized by the LLM. If a prompt scores lower than 0.12, it likely generates a stereotypical output and is optimized by the LLM. This threshold can also be altered with a slider to determine the amount of intervention of the system depending on the use case (Figure 7).

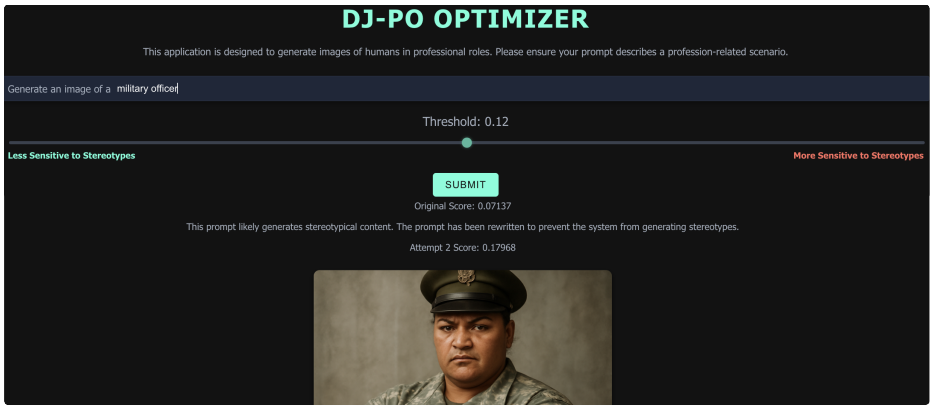

Fig. 7. Frontend DJ-PO Prototype.

To evaluate the effectiveness of the DJ-PO framework, we again asked 20 people to rank 10 sets of 6 images from most to least stereotypical. This time, however, half of the images were generated with the DJ-PO prototype, and the other half were generated by directly calling the T2I model (gpt-image-1). A set of 20 prompts used to generate images of humans in occupations were randomly generated by GPT-4o. Two prompts were not optimized by the DJ-PO prototype, as they did not meet the threshold value of 0.12.

To validate the performance of the DJ-PO prototype, two statistical tests were conducted. First, the assumption of equal variances was tested using Bartlett's test [30]. Since this assumption was violated, Welch's t-test was applied, as it is more robust under unequal variances and appropriate for comparing two groups [31]. The results of these statistical analyses are discussed in Section 4.2.

4 Results

4.1 Reward Model Results

Throughout training, performance of the reward model steadily improved as measured by Spearman's ρ, increasing from 0.0637 in the first epoch to 0.4915 by the final epoch. This indicates a moderate-to-strong positive correlation between the model's predicted prompt rankings and the training data [32]. The result suggests the model effectively learned to align with human judgment.

4.2 DJ-PO Prototype Results

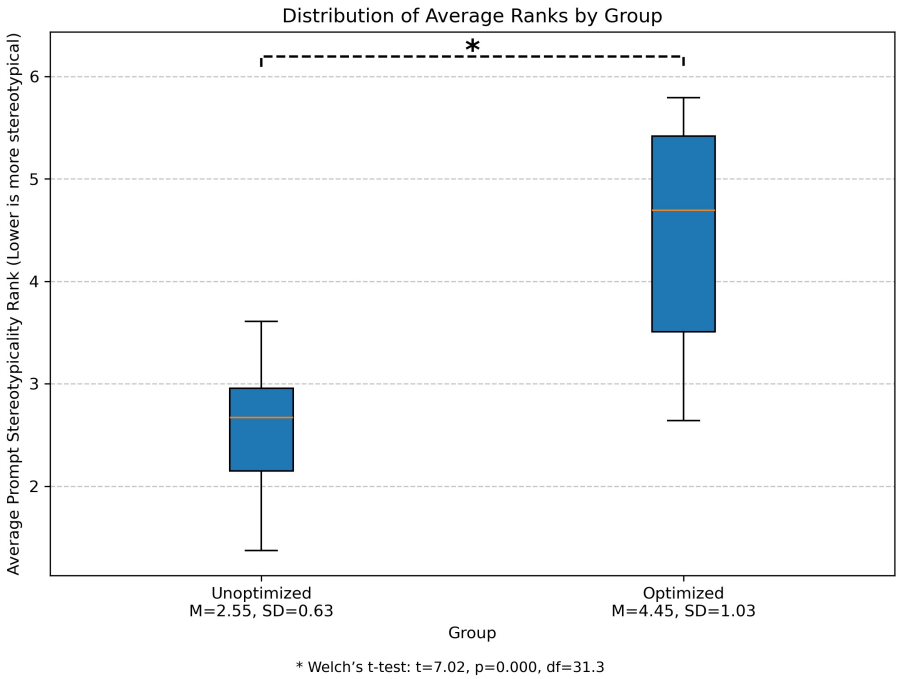

Fig. 8. Welch's T-test DJ-PO Prototype evaluation boxplot.

Bartlett's test yielded a p-value of 0.0355, indicating a significant difference in variances between the optimized and unoptimized prompt groups, thus violating the assumption of homogeneity. Therefore, a Welch's t-test was applied to compare stereotypicality scores. The Welch's t-test (Figure 8) revealed a statistically significant difference in stereotypicality scores between images generated using the DJ-PO optimizer and those generated using the traditional method, $t(31.34) = 7.02$, $p < .001$ [30]. This result indicates that the observed difference is highly unlikely to be due to random variation. Human annotators ($n = 20$, 10 rankings per annotator) consistently rated images produced by the DJ-PO framework as less stereotypical, providing strong quantitative support for the conclusion that prompt optimization via DJ-PO effectively mitigates stereotypical bias in image generation. Figure 9 shows examples of optimized and unoptimized images using the same base prompts. Optimized images show a more diverse representation of the occupational role.

Fig. 9. Examples of Optimized vs Unoptimized images.

5 Conclusion and Discussion

In today's AI landscape, generative models risk reinforcing or amplifying occupational stereotypes, making interventions to promote diversity an urgent challenge. This study introduced the DJ-PO prototype: a framework that combines reward models and LLMs to optimize prompts for reducing stereotypical outputs in text-to-image generation. By working at the prompt level, DJ-PO avoids the high computational costs of image evaluation while still embedding human values such as inclusivity and diversity into the generative process.

Automated diversity injection by randomly adding identity terms to prompts can cause historically inaccurate depictions. To avoid this, DJ-PO introduces only a controlled diversity context into the LLM optimizer, allowing for more context-aware adjustments rather than automatic enforcement. Additionally, it preserves user agency by letting users adjust how strongly stereotype mitigation is applied via a tunable threshold.

The results show that even with a relatively small dataset and without access to model internals, prompt-level optimization offers a lightweight, scalable, and broadly applicable strategy for mitigating bias in T2I systems. An important advantage of this approach is that it requires relatively little effort, resources, or computational power, making the methodology repeatable and accessible. DJ-PO demonstrates that prompt optimization guided by human feedback is a promising but context-limited approach to

reducing stereotypes in generative AI. It shows that practical, resource-efficient interventions can give creative professionals and hobbyists more control over representational diversity in AI-generated content.

While the DJ-PO prototype demonstrates promising results, it is important to acknowledge certain limitations. The technical results serve more as general indicators of system performance rather than conclusive proof of stereotype elimination. Despite a statistically significant difference between the two groups, the optimized prompts may have only produced marginally less stereotypical images than their non-optimized counterparts, as all images were AI-generated and ranked comparatively.

Even though the annotators had varied ethnic backgrounds, the feedback was collected exclusively in the Netherlands, meaning the results are shaped by Western-centric perspectives. The average annotator age of 23.7 further suggests a bias toward the perspectives of younger individuals, shaping a particular interpretation of what is considered stereotypical. This limitation highlights that the system currently reflects a narrow cultural lens rather than universally representative values. Therefore, the contribution of this study is best understood as a proof of concept. For DJ-PO to generalize more broadly, it will be essential to gather annotator feedback from more diverse and clearly defined target groups.

Since LLMs themselves carry biases, prompt optimization can unintentionally replace one stereotype with another. In testing, the DJ-PO prototype's LLM (GPT-4o) often defaulted to Buddhism when asked to include a religion different from the one typically associated with an occupation. This indicates substitution rather than genuine diversification and underscores the need for careful prompt engineering and ongoing evaluation, as systems intended to reduce stereotypes may inadvertently introduce new ones. Additionally, the current framework's focus on seven diversity dimensions (Appendix 1) is a strong starting point, but it still simplifies the multidimensional nature of human diversity and risks introducing new stereotypes.

DJ-PO specifically targets occupational imagery, where deviating from historical norms can be normatively justified depending on the societal context. At the same time, users retain the ability to lower the diversity threshold when historical accuracy is desired. The current diversity threshold is based on 240 prompts and may not generalize well, highlighting the need to include real user-made prompts in future evaluations to improve external validity.

Overfitting was observed in syntactic patterns of prompts: "Generate an image of a male CEO" was scored as more diverse than "Generate an image of a CEO," indicating possible overcorrection or bias toward specific wordings. Ideally, a system trained on more and higher quality data would recognize that the combination of the words "male" and "CEO" should result in a lower diversity score.

Furthermore, prompt optimization currently occurs only in the backend. Making optimized prompts visible to users would increase transparency and provide insight into how the system modifies the prompt provided by the user, helping users understand and adjust the degree of intervention more consciously.

The DJ-PO prototype offers media organizations a plug-and-play solution to reduce stereotypes in generative image systems. More generally, its methodology can embed

subjective human values such as beauty, humor, or emotional appeal into reward models. Relying on human rankings, DJ-PO is especially suited for tasks involving subjective interpretation: annotators can evaluate outputs on dimensions like engagement, aesthetics, or relatability, allowing adaptation beyond stereotype reduction. The DJ-PO approach is not limited to occupations: the same methodology could be applied to other target groups, different stereotype dimensions, or even entirely different categories of human representation, underscoring the versatility and transferability of the framework. Importantly, different audiences hold different value systems; for example, a model trained on European aesthetics may behave differently from one trained on Asian values, highlighting the role of cultural context in aligning generative models with human preferences.

Future research should prioritize increasing data quantity to avoid reliance on synthetic data collection methods. Relying on synthetic data likely reduced nuance and masked more subtle signals of stereotypicality. Additionally, although increasing the number of trainable layers did not initially improve performance, deeper models may prove more effective at learning fine-grained distinctions in larger, non-synthetic datasets. Future work should also compare stereotypes present in AI-generated images with those found in real-world imagery to better understand how T2I models reflect or amplify existing societal biases.

Another future research direction involves improving contextual optimization and ensuring the model can better distinguish when diversity is appropriate versus when it distorts the intended meaning of a prompt. Building upon this, another valuable direction would be testing whether the DJ-PO prototype generalizes to other or newer T2I models beyond the one evaluated in this study. Evaluating its effectiveness across different generative backends could strengthen its applicability and robustness.

Future iterations of DJ-PO might also benefit from leveraging recent advances in memory-augmented architectures or in-context learning. By allowing the LLM to retain or recall previous prompt optimization strategies, the system could learn to diversify prompts more effectively over time, especially when tasked with generating content across a broader range of identities or cultural contexts.

Acknowledgments. We would like to thank Javier Sancho Rodriguez and the team at Monks for their valuable practical feedback, which helped keep this research grounded and relevant to real-world applications. This research was made possible in part by the SPRONG project Responsible Applied AI, grant number SPR.ALG.01.024, funded by the Dutch Government through Regieorgaan SIA. We acknowledge the use of LLM's for assistance in language polishing.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nicoletti, L., Bass, D.: Humans Are Biased. Generative AI Is Even Worse, https://www.bloomberg.com/graphics/2023-generative-ai-bias/, (2024).

2. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In: ACM Conference on Fairness, Accountability, and Transparency. pp. 1493–1504. ACM, Chicago IL USA (2023). https://doi.org/10.1145/3593013.3594095.
3. He, Y., Robey, A., Murata, N., Jiang, Y., Williams, J., Pappas, G., Hassani, H., Mitsufuji, Y., Salakhutdinov, R., Kolter, J.Z.: Automated Black-box Prompt Engineering for Personalized Text-to-Image Generation. (2024).
4. Marx, D., Ko, S.J.: Stereotypes and Prejudice. In: Oxford Research Encyclopedia of Psychology (2019). https://doi.org/10.1093/acrefore/9780190236557.013.307.
5. Macrae, C.N., Stangor, C., Hewstone, M.: Stereotypes and Stereotyping. Guilford Press (1996).
6. Dharani, B.: Fluidity of social identities: implications for applying intersectionality. Equal. Divers. Incl. Int. J. (2024). https://doi.org/10.1108/EDI-06-2022-0151.
7. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623. ACM, Virtual Event Canada (2021). https://doi.org/10.1145/3442188.3445922.
8. Hudson, S.T.J., Myer, A., Berney, E.C.: Stereotyping, prejudice, and discrimination at the intersection of race and gender: An intersectional theory primer. Soc. Personal. Psychol. Compass. 18, e12939 (2024). https://doi.org/10.1111/spc3.12939.
9. Ghosh, S., Venkit, P.N., Gautam, S., Wilson, S., Caliskan, A.: Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach, http://arxiv.org/abs/2407.14779, (2024). https://doi.org/10.48550/arXiv.2407.14779.
10. Mañas, O., Astolfi, P., Hall, M., Ross, C., Urbanek, J., Williams, A., Agrawal, A., Romero-Soriano, A., Drozdzal, M.: Improving Text-to-Image Consistency via Automatic Prompt Optimization, http://arxiv.org/abs/2403.17804, (2024). https://doi.org/10.48550/arXiv.2403.17804.
11. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., Sui, Z.: A Survey on In-context Learning, http://arxiv.org/abs/2301.00234, (2024). https://doi.org/10.48550/arXiv.2301.00234.
12. Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.-R., Yang, Q.: Dynamic Prompt Optimizing for Text-to-Image Generation. (2024).
13. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., Kaplan, J.: Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, http://arxiv.org/abs/2204.05862, (2022). https://doi.org/10.48550/arXiv.2204.05862.
14. Kaufmann, T., Weng, P., Bengs, V., Hüllermeier, E.: A Survey of Reinforcement Learning from Human Feedback, http://arxiv.org/abs/2312.14925, (2024). https://doi.org/10.48550/arXiv.2312.14925.
15. Veiligheid, M. van J. en: Wat zegt de wet over discriminatie? - Mensenrechten voor jou - College voor de Rechten van de Mens, https://www.mensenrechten.nl/mensenrechten-voor-jou/discriminatie-en-gelijke-behandeling/wat-zegt-de-wet-over-discriminatie-en-gelijke-behandeling, last accessed 2025/04/17.

16. Tiggemann, M., Slater, A.: Thin ideals in music television: A source of social comparison and body dissatisfaction. Int. J. Eat. Disord. 35, 48–58 (2004). https://doi.org/10.1002/eat.10214.
17. Neumark, D.: The Age Discrimination in Employment Act and the Challenge of Population Aging. Presented at the September (2008). https://doi.org/10.1177/0164027508324640.
18. McMahon, B.T.: An overview of workplace discrimination and disability. (2012). https://doi.org/10.3233/JVR-2012-0588.
19. Lee, S.: A Comprehensive Guide to Bartlett's Test for Variance Analysis, https://www.numberanalytics.com/blog/bartletts-test-variance-guide, last accessed 2025/06/07.
20. Bobbitt, Z.: How to Check ANOVA Assumptions, https://www.statology.org/anova-assumptions/, last accessed 2025/09/08.
21. Bobbitt, Z.: A Guide to Using Post Hoc Tests with ANOVA, https://www.statology.org/anova-post-hoc-tests/, last accessed 2025/09/08.
22. Qi, J., Du, J., Siniscalchi, S.M., Ma, X., Lee, C.-H.: On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression. IEEE Signal Process. Lett. 27, 1485–1489 (2020). https://doi.org/10.1109/LSP.2020.3016837.
23. Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th international conference on Machine learning. pp. 129–136. ACM, Corvalis Oregon USA (2007). https://doi.org/10.1145/1273496.1273513.
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network, http://arxiv.org/abs/1503.02531, (2015). https://doi.org/10.48550/arXiv.1503.02531.
25. Zhuang, Z., Liu, M., Cutkosky, A., Orabona, F.: Understanding AdamW through Proximal Methods and Scale-Freeness. Aug 2022.
26. Mallick, U.: Cosine Learning Rate Schedulers in PyTorch, https://medium.com/@utkrisht14/cosine-learning-rate-schedulers-in-pytorch-486d8717d541, last accessed 2025/09/02.
27. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 04, 2019, *arXiv*: arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101.
28. J. Zar, "Spearman Rank Correlation," in *Encycl Biostat*, vol. 5, 2005. doi: 10.1002/0470011815.b2a15150.
29. Drapala, J.: Kernel Density Estimation explained step by step, https://towardsdatascience.com/kernel-density-estimation-explained-step-by-step-7cc5b5bc4517/, last accessed 2025/08/25.
30. Field, A.: Discovering Statistics Using IBM SPSS Statistics (6th ed.). SAGE (2024).
31. Difference between t-test and ANOVA, https://www.geeksforgeeks.org/difference-between-t-test-and-anova/, last accessed 2025/06/17.
32. J. Zar, "Spearman Rank Correlation," in *Encycl Biostat*, vol. 5, 2005. doi: 10.1002/0470011815.b2a15150.

6 Appendix

6.1 The seven Diversification aspects

"With a different religion or belief than usually associated with the job",
"With a different race or ethnic origin than usually associated with the job",

"With a different gender identity or expression (e.g., male, female, transgender, gender fluid and more...) than usually associated with the job",
"With a different nationality than usually associated with the job (e.g., a specific country in Europe, Asia, Africa, North America, South America, Antarctica, or Australia)",
"With a different ability or disability status than usually associated with the job",
"With a different age or generational identity than usually associated with the job",
"With a different body type than usually associated with the job"