Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models

Eddie L. Ungless
School of Informatics
University of Edinburgh
Scotland

Björn Ross
School of Informatics
University of Edinburgh
Scotland
b.ross@ed.ac.uk

Anne Lauscher
Data Science Group
University of Hamburg
Germany

Abstract

Warning: many of the images displayed in this paper are offensive and contain implied nudity. They are intended to illustrate potential harms.

Cutting-edge image generation has been praised for producing high-quality images, suggesting a ubiquitous future in a variety of applications. However, initial studies have pointed to the potential for harm due to predictive bias, reflecting and potentially reinforcing cultural stereotypes. In this work, we are the first to investigate how multimodal models handle diverse gender identities. Concretely, we conduct a thorough analysis in which we compare the output of three image generation models for prompts containing cisgender vs. non-cisgender identity terms. Our findings demonstrate that certain non-cisgender identities are consistently (mis)represented as less human, more stereotyped and more sexualised. We complement our experimental analysis with (a) a survey among non-cisgender individuals and (b) a series of interviews, to establish which harms affected individuals anticipate, and how they would like to be represented. We find respondents are particularly concerned about misrepresentation, and the potential to drive harmful behaviours and beliefs. Simple heuristics to limit offensive content are widely rejected, and instead respondents call for community involvement, curated training data and the ability to customise. These improvements could pave the way for a future where change is led by the affected community, and technology is used to positively *"[portray] queerness in ways that we haven't even thought of"* rather than reproducing stale, offensive stereotypes.

1 Introduction

Summer 2022 saw the publicly accessible DALL·E mini text-to-image model go viral (Hughes, 2022). Users enjoyed creating and sharing digital art, with some 50,000 images being produced a day (Knight,

Figure 1: Four images generated by Stable Diffusion model in response to *"Transgender women"*. The black square indicates the model did not produce an output due to risk of NSFW content.

2022). Very quickly, a form of *"everyday algorithmic auditing"* began (Shen et al., 2021), whereby users of the model shared potentially harmful images produced in response to neutral prompts[1]. Some of the generated images reflected human stereotypes such as the association between the roles of CEO and programmer, and white men – a finding corroborated by recent research (Bianchi et al., 2022; Bansal et al., 2022; Cho et al., 2022).

Text-to-image models reflect social biases in their output, just as word embeddings and neural language models have been shown to capture related gender and racial stereotypes (Bolukbasi et al., 2016; Guo and Caliskan, 2021; Sheng et al., 2019a). Biased text-to-image models may result

---

[1] https://twitter.com/jose_falanga/status/1537953980633911297, https://twitter.com/ScientistRik/status/1553151218050125826, https://twitter.com/NannaInie/status/1536276032319279106

both in representational harms, where harm occurs due to how a particular sociodemographic is represented, and allocational harms, relating to the allocation of resources to the sociodemographic such as access to job opportunities and the ability to use a service (Barocas et al., 2017).

Our own "everyday auditing" of DALL·E mini revealed potentially offensive content produced in response to non-cisgender[2] identity terms: images were often cartoonish and figures were rendered using colours from associated flags, adding to the lack of realism, which could reinforce the belief that such identities "aren't real" (Valentine, 2016; Minkin and Brown, 2021). Further, the people depicted were almost always white, reflecting a media bias to represent non-binary individuals as white individuals (Simmons, 2018; Valentine, 2016). We build on this with a **systematic annotation study of content produced by three text-to-image models** in response to prompts containing different gender identities, such as the ones given in Figure 1. Identifying whether the model produces harmful content in response to non-cisgender identities allows us to caution the research community and public when developing and using these models.

In order to expand beyond our own preconceptions, we also conduct a **survey of non-cisgender individuals**, asking them to identify potential harms of the model. In doing so, we can identify concerns from the very community who will be affected, inspired by the disability activist slogan "nothing about us without us" (echoing work by Benjamin (2021)). Finally, beyond identifying harms, we explore the communities' desired output from these models with regards to representing their identities, through a series of **interviews**.

Contributions. Our main contributions are as follows: (1) We are the first to present a thorough manual analysis of how text-to-image models currently handle gender identities in different application contexts, and the potential harms, to highlight the caveats of these models. (2) We provide recommendations for how models should be shaped in future based on how the community would like to be represented. Our findings will provide guidance to those developing the models as to how the affected community would like for these issues to be

---

[2]We use "non-cisgender" as an umbrella term for those who do not identify as cisgender men or cisgender women, including trans, non-binary, gender non-conforming, agender, third gender, latinx and Two-spirit identities.

resolved. Providing this kind of insight is crucial to ensuring the voices of those who are marginalised are heard and used to lead development, rather than work being guided by the intuitions of those who are not impacted by such harm.

2 Related Work

We survey the literature relating to (gender) identity inclusion in NLP and the recently emerging area of bias analysis in image generation.

Identity-Inclusive NLP. Existing work on non-cisgender identities and machine learning is sparse (e.g., Dev et al., 2021; Cao and Daumé III, 2020; Lauscher et al., 2022). However recently, there have been a couple of works dealing with gender-neutral pronouns (e.g., Brandl et al., 2022; Qian et al., 2022). As such, work by Lauscher et al. (2022) explores the diversity of gender pronouns and presents five *desiderata* for how language models should handle (gender-neutral) pronouns. In a similar vein, we explore potential solutions for how text-to-image models should handle non-cisgender identities. Brandl et al. (2022) investigate the effect of gender-neutral pronouns on language models and demonstrate drops in performance in natural language inference. As a potential solution, Qian et al. (2022) propose a perturber model for augmenting data sets which they train on texts that have been rewritten in a gender-neutral way. Most relevant to our approach, Dev et al. (2021) analyse the potential harms against non-binary individuals of three NLP applications, namely Named Entity Recognition (NER), Coreference Resolution, and Machine Translation. They survey non-binary individuals with AI experience to identify possible harms for these tasks, and in different domains. They additionally analyse the potential for erasure and misgendering due to use of GloVe or BERT embeddings. We extend their work by analysing potential harms of text-to-image models, and additionally consider how the community would like to be represented by these models.

Bias Analysis in Image Generation. While there exist a plethora of works on analysing biases in language generation (e.g., Sheng et al., 2019b; Yeo and Chen, 2020; Barikeri et al., 2021, *inter alia*), work on bias in image generation is still relative sparse (e.g., Bianchi et al., 2022; Bansal et al., 2022; Cho et al., 2022). As one of the earliest works, Salminen et al. (2020) found that fa-

cial images generated by StyleGAN (at the time a state-of-the-art image generator) skewed towards young white women. In a similar vein, Struppek et al. (2022) investigated cultural biases. Similar to us, they focus on DALL·E 2 and Stable Diffusion. Cho et al. (2022) probe these models for social stereotypes related to gender and skin colour. Most recently, Bianchi et al. (2022) also explore the topic of bias in text-to-image model outputs, with a focus on stereotyping. In the supplementary material, they also present images generated using the term "non-binary", but don't explore the issue more thoroughly. In our own work, we focus not only on stereotypes, but also the quality of images produced for diverse gender identities and provide an empirical analysis of the issues. Identifying these kinds of biases in text-to-image models allows for more targeted mitigation strategies.

3 Analysis of Generations

We investigate how models currently handle gender identities. We insert gender identity terms into template prompts, generate images using three state-of-the-art models and annotate image features such as photorealism and implied nudity to compare cisgender and various non-cisgender identities.

3.1 Prompt Creation

We used five neutral templates (with little inherent meaning) and five templates designed to represent possible commercial use of the models, given in Table 1. All prompts are in English. This small number of templates allowed us to focus on variation across a large number of identities (which we prioritise over exploring linguistic diversity). The "commercial" templates were taken from Conceptual Captions, a dataset of images and HTML-alt text (Sharma et al., 2018). We manually selected five captions from the unlabelled training data that included *person, woman* or *man*, then replaced this with one of our identity phrases. We use these real world captions to improve the ecological validity of our analyses (that is to say, how well the experimental findings relate to the real world). We selected captions that relate to commercial use cases identified in the DALL·E 2 documentation[3].

We identified ten words relating to trans status, namely *cisgender, latinx, two-spirit, transgender, trans, enby, nonbinary, gender non-conforming,*

genderqueer and *queer*; and combined where appropriate with person terms (*woman, man, person, women, men, people*) and pronouns from the list *his, her, their, xyr, its*. Our choice of terms was based on a recent "Gender Census", with the addition of *two-spirit, latinx* to expand our focus to identities used exclusively by people of colour. Term selection and use is explained in Appendix A. Whilst some of these identity terms have multiple meanings, for example *queer, latinx*, we wanted to be inclusive in our choice of terms, acknowledging that language use can be "fuzzy".

We also include examples where trans status is not specified, but cisgender will be "assumed" (in the sense that the training data will almost exclusively include examples where trans status is not specified but the individuals depicted are cisgender), as this is the norm (Bucholtz and Hall, 2004; DePalma and Atkinson, 2006). This allows us to explore how the model handles implicit norms (where trans status is not given but cisgender will be assumed) and explicit norms (where cisgender is stated), and also allows us to control for word length (though how the models handle tokenisation will impact how the input is processed). We detail how we combine these terms in Appendix A. The large number of possible trans status, person and pronoun combinations gave 231 prompts when combined with our 10 templates.

We used sentence case but no final punctuation in our prompts, to match typical prompt usage observed on Twitter and in prompt guides [4].

3.2 Image generation

We generated four images for every prompt. Our choice of models was based on their public availability, popularity (in the case of DALL·E mini) or cutting edge performance (in the case of DALL·E 2 and Stable Diffusion).

DALL·E mini We used the dalle-mini/dalle-mega model (henceforth `dall-e mini`) (Dayma et al., 2021). The public facing DALL·E mini app incorporates both "DALL·E Mini" and "DALL·E Mega" models. Images were generated using an adapted version of a Python notebook[5] (adapted to run as a script using the chosen `dalle-mini`

---

[3] https://github.com/openai/dalle-2-preview/blob/main/system-card.md

[4] https://dallery.gallery/the-dalle-2-prompt-book/

[5] https://colab.research.google.com/github/borisdayma/dalle-mini/blob/main/tools/inference/inference_pipeline.ipynb

Neutral Templates	Commercial Templates
The *(trans status) person*	Vector illustration of *(trans status) person* isolated on a white background
A *(trans status) person*	Portrait of a smiling *(trans status) person* stroking *pronoun* dog lying on couch
(trans status) person	A demonstration of a group of *(trans status) people* practicing their rights
(trans status term) people	*(trans status) people* tour and enjoy the public park during summer
(trans status)	*(trans status) person* at a corporate event

Table 1: Templates indicating where trans status phrases, person and pronoun terms are included. (Parentheses) indicate optional elements. *Person* is replaced with *man, woman* where appropriate. *People* is replaced with *men, women* where appropriate. Pronoun is replaced with *his, her, their, xyr, its* where appropriate.

model, and to generate four images for each prompt). Images were produced in <2 GPU hours.

DALL·E 2 For generating images with DALL·E 2, we resorted to OpenAI's Python package[6] and queried the paid image generation API with our prompts (resolution set to 256x256 pixels).

Stable Diffusion We used the most popular Stable Diffusion text-to-image model on Hugging Face, namely `stable-diffusion-v1-5` (Rombach et al., 2022), henceforth Stable Diffusion, with default parameters, creating four images/ prompt. Images were produced in <2 GPU hours.

3.3 Annotation Procedure

We recruited six annotators from our institutions that (a) were all familiar with the concept of AI-based image generation, (b) were proficient speakers of the English language, (c) represented relatively diverse cultural, and gender backgrounds, and (d) demonstrated great interest in helping to make AI more inclusive. Annotators were based in Europe. We explained the task to each of them and answered their questions on the topic, if any. Annotators were aware they may see offensive and NSFW material. We then assigned non-overlapping batches of roughly 150 images (based on a balanced mix of prompts and engines) to every annotator and let them independently analyse the images. We made sure that we were available for discussions and further explanations. Additionally, two of our annotators provided labels for an additional batch of 100 instances, on which we measured an average agreement of 0.8 Krippendorff's α across all questions with the lowest score on the question whether annotators see a flag (0.56) and the highest on whether there is an individual

present (1.00). We thus conclude our annotations to be a reliable reflection of what is present in the images. The total number of annotated instances is 984.[7]

Annotators were asked to indicate:

- Level of photorealism
- Whether an individual is present, and if so:
  - How many individuals are visible?
  - Are facial features mostly visible?
  - Is anyone non-white?
  - Is there (implied) nudity of torso or crotch?
- Are there text or symbols?
- Are there flags in the image?
- Is the image relevant to the template? (for the "commercial" templates only)

Our choice of features to annotate is based on research into dehumanisation, which we detail in Appendix A, and results of our initial audit of DALL·E mini. We also depict the interface in Appendix A.

3.4 Results: Qualitative Observations

We first conduct a qualitative analysis of the results.

Stereotypes All three models struggled with the identity Two-spirit, often rendering pairs of individuals. Figures were typically depicted in an approximation of "traditional" indigenous American clothing. Images were often cartoon-like or very abstract. Figure 6 in Appendix A illustrates these points. Notably, one of our interview participants expressed a concern that these models would fail to portray indigenous non-cisgender individuals

---

[6] https://github.com/openai/ openai-python

[7] Instead of 1,000, because some annotators accidentally skipped some images.

as "normal", showing them in religious garb; this concern seems to have been realised.

Often, images of transgender people contained a combination of features typically regarded as "feminine" and "masculine" i.e. a beard and long hair. Whilst this is absolutely an expression of gender that exists and is to be celebrated, some transgender people may feel this is an inaccurate representation (and indeed our interviewees expressed concern about the inappropriate combining of typically gendered features, see Section C). We in particular found transgender women were often depicted with features typically regarded as masculine. Figure 7 in Appendix A illustrates this point. Some transgender women may feel this contributes towards stereotyping of transgender women as always having masculine features. We conclude that **models propagate ethnic and trans status stereotypes.**

Smut We found the models produced pornographic images despite attempts to mitigate this by developers. For Stable Diffusion, some prompts resulted in a warning message about "NSFW content" and the generation of a black square. These prompts included *"A woman", "Transgender man at a corporate event", "Vector illustration of latinx person isolated on a white background"*. The most common identity affected was *"transgender"*. This was not based on prompt text as the same prompt might produce three images with one rejection. Despite this safety step, the model produced a number of pornographic images including graphic images of genitalia. DALL·E 2 "refused" to generate an image for a number of prompts derived from the template *"Portrait of a smiling <identity phrase> stroking <pronoun> dog lying on couch"* and the identity terms *cisgender, trans* and *transgender*, stating "Your prompt may contain text that is not allowed by our safety system". We believe the word "stroking" combined with a trans status term may have triggered this warning, although some combinations were allowed, as were unmarked identities (*man, woman, person*). We thus conclude that **prompt blocking and NSFW warning features are likely to contribute to the erasure of non-cis identities and often do not prevent the generation of harmful output.**

3.5 Results: Annotation Task

We show some of the results of our analysis in Figures 2a–2d. The average degree of photorealism varies slightly among images generated with prompts containing different identity phrases (Figure 2a). Images for latinx identity phrases achieve the highest average score with 2.8, followed by phrases commonly associated with cisgender identities (e.g., *man, woman, etc.*) with 2.7. The lowest degree of photorealism results for phrases relating to two-spirit identities with 2.2. There is a large variation in the proportion of images containing symbols and text (Figure 2c) or flags (Figure 2d). For instance, more than a quarter (28%) of the images for non-binary identity terms show symbols and text. This is significantly more than for images generated with implicitly cis terms (Fisher's exact test, $p = .038$). Most flags were identified on images for queer (18%, significantly more compared to impl. cis, $p < .001$), latinx (15%), and trans (12%) identity phrases. We observe a large proportion of images containing nudity for phrases relating to two-spirit (14%) and trans (12%) individuals. The differences between images generated with implicitly cis vs. two-spirit ($p = .009$) and trans ($p = .016$) identity terms are also statistically significant. We further note a high amount for phrases explicitly conveying cis-identity (8%) possibly triggered by the token *"gender"*. Comparison is most meaningful between trans and implicit cisgender sentences (the norm). Figures 4 and 5 in the Appendix illustrate this point: there is a stark difference in the amount of nudity in response to two prompts that differ only by the word "transgender".

We observe a lack of ethnic diversity in the images: the majority of images contain no non-white individuals. Figures 1, 5 and 7 in the Appendix illustrate this point. The models reflect the (Western) norm of whiteness. In sum, there is high output variation depending on the identity phrase in the prompt, which is likely to lead to a lower degree of photorealism and potentially harmful generations (i.e., nudity, stereotypes).

4 Survey of Non-Cisgender People's Expectations

We conducted a survey of English-speaking non-cisgender individuals to investigate potential harms. We also asked respondents for their satisfaction with a number of heuristic solutions, and optionally to provide their own solutions to the harms.

(a) Photorealism

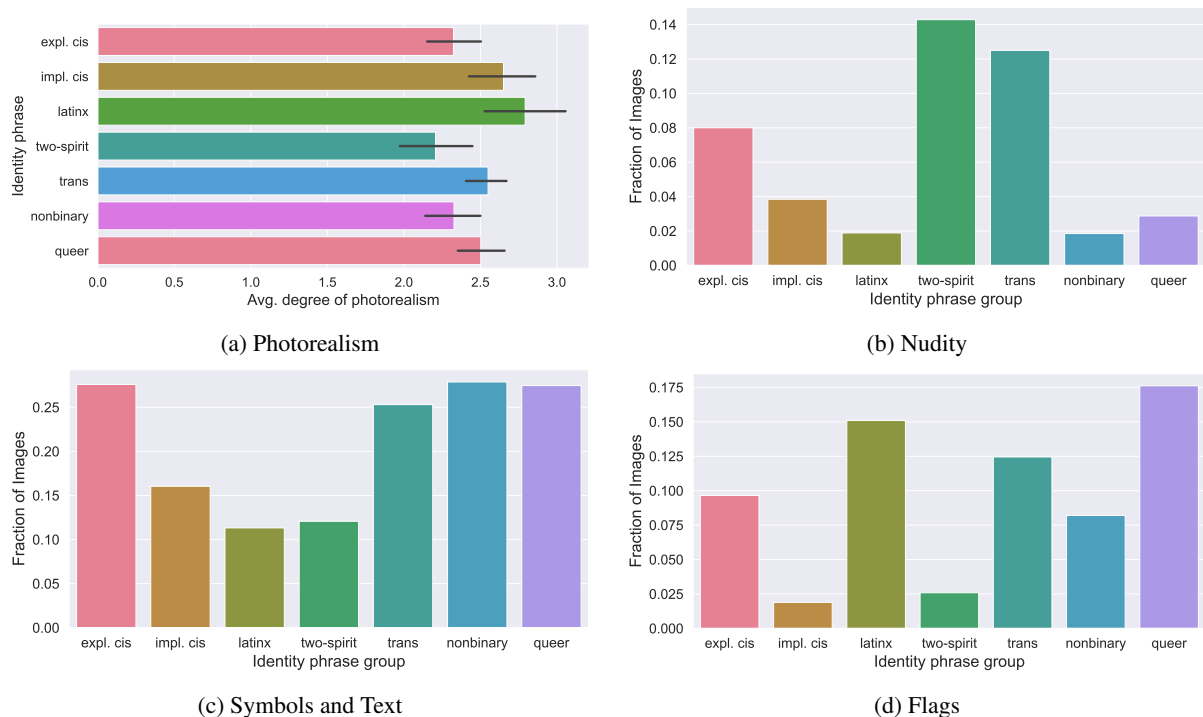(b) Nudity

(c) Symbols and Text

(d) Flags

Figure 2: Results of our image analysis. We show (a) the average degree of photorealism, (b) the fraction of images with nudity content, (c) the fraction containing symbols, and (c) the fraction with flags per identity phrase group (*explicitly cis*, *nonbinary*, *queer*, *latinx*, *comm. cis*, *trans*, and *two-spirits*) aggregated over all three engines.

4.1 Methodology

4.1.1 Participants

We recruited participants through posts on social media and the Queer in AI community group. Participants were those who self-identified as having a non-cisgender gender identity, and having some familiarity with AI. We hope that our focus on those with some familiarity with AI will allow us to explore the topic in depth without use of leading questions – participants can draw on their own experience of issues that have arisen in their work, and their familiarity with ML techniques will provide them with foresight as to the kinds of problems that might arise. In this we are following the success of Dev et al. (2021) in their study on harms of gender exclusivity in language technologies.

4.1.2 Design

Our questions around harms and norms are framed around the potential (commercial) use cases for text-to-image models. We provide examples from the DALL·E 2 documentation[8]. We are not interested solely in the DALL·E family of models, but felt that the proposed usage contexts would provide

a useful starting point for discussions. Participants can relate their answers to potential real-world use cases, providing their own suggested uses also.

4.1.3 Procedure

After giving consent, participants were asked optional demographic questions. The list of questions and answer options are largely taken from Dev et al. (2021), with some excluded for brevity. We asked about gender identity, sexuality, trans status, pronoun use, ethnicity, native languages and experience with AI. We provide full details including a breakdown of answers in Appendix B.

Participants were then given a brief description of text-to-image models, including an example output from the Craiyon[9] model. We outlined how such models were trained (here participants' existing familiarity with AI was crucial to keep descriptions brief). We explained we were interested in exploring these models' potential for harm.

We then presented them with a quote from the DALL·E 2 documentation where they outline potential commercial use cases, explaining our choice of providing these use cases. We asked participants if they could foresee harm occurring through use of this technology in these use cases, and in which use

---

[8] https://github.com/openai/
dalle-2-preview/blob/main/system-card.md

[9] https://www.craiyon.com/

cases. We asked them to rate the potential severity of these harms. This framing could be argued to prime our respondents to agree that harm was likely, but our results indicate that respondents were willing to reject this premise. We then asked them to give an example scenario where harm might occur.

We then presented seven proposed solutions for how models should handle non-cisgender identities and asked users to rate how satisfactory they found each solution. They could optionally provide potential harms and benefits for each solution, and their own proposed solution. Participants were then asked if they had anything to add, then debriefed.

4.2 Survey Results and Discussion

4.2.1 Demographic Information

We had 35 respondents to our survey. Full details are reported in Appendix B. Respondents' ages ranged from 19 to 57, suggesting we were able to capture views from an age-diverse group. The most common gender identity was nonbinary, with 71% of respondents identifying as such (potentially alongside other identities). 85% of our respondents identify as trans, suggesting our avoidance of the terms trans or transgender in our recruitment allowed us to appeal to a wider spectrum of marginalised non-cisgender people.

Only three respondents identified as Black, Latinx and/or Indigenous; similarly, three identified as a person of colour. The vast majority (30) of our participants identified as white/Caucasian. Almost all our respondents (34) currently reside in North America, Europe or Australia, meaning our findings largely reflect a white Western perspective.

All participants rated themselves as having some familiarity with AI, through their education, career and/or personal interests.

4.2.2 Potential for Harm

The overwhelming majority of respondents felt that there was potential for harm, on average rating the severity as moderate. Contexts where a clear majority of users felt harm would occur were in marketing, education and art/creativity, and this was reflected in written responses also. We coded their written responses to the task asking for specific scenario(s) where harm might occur using a deductive-inductive approach. We wished to investigate the presence of allocational and representational harms, and references to the specific contexts of use, but we also developed codes based on the responses.

Representational harms far outnumbered allocational harms suggesting these were most salient to the community. Respondents spoke of their concerns about intentional misuse to create offensive content or harmful technologies. The potential impact on real-world behaviours and beliefs was a common theme, for example the reinforcement of prejudices or the creation of narrow beauty standards. Many respondents made explicit reference to the training data being the source of harm, reflecting the technical experience of our respondents. Details of our analysis are in Appendix B.

4.2.3 Proposed Solutions

We proposed seven solutions that relied on simple heuristics to prevent harmful content being produced, developed through our own experience of heuristics used by existing models, and through casual discussion with colleagues and community members in response to the harmful images produced during the annotation task. The heuristics we proposed were as follows:

- The model generates an image based on the text (no change to current behaviour).

- The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text.

- The model generates an image based on the text but includes a warning that the output might be offensive.

- The model ignores all gender identity terms in the text input and generates an image based on the rest of the text.

- The model is trained on additional images containing non-cisgender individuals, so it better learns to generate images of non-cisgender people.

- The model effectively ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, but a flag or pin or symbol is used to indicate gender diversity.

- The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, with a warning that to avoid harmful misrepresentation the model ignores non-cisgender identity terms.

The "solution" to change nothing was considered fairly unsatisfactory, with respondents noting concerns about stereotyping, although some respondents considered this their preferred outcome. The proposed heuristic solutions such as ignoring non-cisgender identities terms (with or without an indication); ignoring all gender identities terms, and including a warning that the output might be offensive, were all deeply unpopular. However, the range of ratings indicated a diversity of opinions – for example, the suggestion to "[include] a warning that the output might be offensive" received a low average rating but the bimodal nature of the results suggests there was a subset of respondents who found this solution to be somewhat satisfactory (see Figure 12 in the Appendix).

By far the most satisfactory solution was to increase the amount of training data. However, respondents expressed concerns about the challenge of collecting representative data, and some were worried about the safety ramifications of gathering a labelled dataset of marginalised individuals. Full analysis of responses related to heuristics can be found in Appendix B.

Respondents were also invited to provide their own solutions for how they would like to models to handle non-cisgender identities. We coded their answers using an inductive approach, and found a number of key themes emerge related to the topic of how respondents wish to be represented, namely the need for representative data; unhappiness with the proposed heuristics; the necessity of wider changes; the need for community involvement; a desired ability to customise images. For example, participants called for "a diverse and representative set of images" in the training data, of queer and other marginalised identities, but also felt that "fixing society generally" may be necessary for technology to not produce harmful content. Our thematic analysis can likewise be found in Appendix B.

5   Interviews

We additionally interviewed four participants who had indicated interest in the survey, selected to engender a diversity of views. We wanted to explore the potential harms in more depth, and in particular we wanted to discuss participants' preferences for how they would like to be represented, which we felt could be challenging to describe in text alone. Just as our survey aimed to expand beyond our preconceived harms, the interviews aimed at expanding beyond our preconceived solutions. Methodology and full analysis of results can be found in Appendix C.

The seven major themes we identified in participants responses were harmful output; being unable to use current technology; rejection of heuristics; need for community input; need for transparency and regulation; desire for authentic representation; the potential for good.

As we found in the survey, participants expressed unhappiness with the heuristic solutions and in particular the idea of appending "warning labels" – they felt this could lead to the community being associated with offensiveness. They were concerned the heuristics would lead to erasure of the community. However, participants were also concerned about unintentional and intentional harms, and many felt they could not use the current technologies. Their concerns included the potential for real world repercussions and even "violent stuff in the long run".

Participants suggested instead greater community involvement at every step, and greater transparency and regulation as the way to ensure more representative output. In addition to involving non-cisgender people at every stage of development, the community could provide feedback on what output they feel is "right for them".

Participants felt that since use of these technologies seems "inevitable", the models must produce authentic representations of humanity: for example, the true global diversity of gender expressions should be captured, including "different expressions of gender in the global south".

Participants spoke of the "potential" to use image generation technologies to imagine queer futures for the community, which "can be... exciting", either through representing themselves in ways more aligned with their internal sense of self, or "[portraying] queerness in ways that we haven't even thought of".

6   Where to go from here?

We identified a great potential for harm through our annotation task and surveys and interviews with community members. Our annotation task revealed dehumanisation, othering, stereotyping and sexualisation of non-cisgender identities. Community members were concerned about misrepresentation, and intentional misuse of the technologies, as well

as the potential for output to negatively influence people's behaviours and beliefs.

Rejection of heuristics   Heuristic solutions to the problem of misrepresentation of non-cisgender individuals were almost universally rejected. Whilst we did not directly ask about this scenario, the Stable Diffusion and DALL·E 2 models' behaviour of refusing to generate potentially NSFW content would likely have been rejected as well by survey and interview respondents who spoke repeatedly of the harms of not being represented or being associated with warning labels. Unfortunately, the association between transgender identities and pornography means images of these communities are likely to be subject to greater censorship.

Curation of training data   Respondents favoured curated training data as a way to improve representation, though they expressed hesitation over whether such a compiled dataset would be safe, and whether it could ever be truly representative. Careful, community led data curation may address some of these concerns, including involvement in creating sensitive labels for images.

Visualising the unseen   Some communities are likely to remain underrepresented in training data for technical or safety reasons, or because the community is small. Models rely on huge amounts of data; novel novel data-efficient strategies that allow for adequate (and potentially customizable) representation of individuals that identify with small communities are needed to address the representation of such communities.

Desire for customisation   The ability to customise images was proposed as a novel solution, which may help to overcome a lack of suitably diverse training data. Whilst this level of customisation is still emerging (OpenAI have recently introduced an Outpainting feature allowing users to generated extensions of a generated image[10]), our survey suggests this is a desirable feature for handling diverse identities appropriately. The lack of ability to customise was mentioned as a potential harm of these models by one respondent. Such customisation would also help with creating more faithful representations of other non-normative identities. Of course, as (Brack et al., 2022) note, a drawback

[10]https://openai.com/blog/dall-e-introducing-outpainting/

would be that such image customisation could also be used to create more harmful content.

Need for community involvement   Respondents felt community involvement would help address some issues, but societal level changes were called for to make meaningful improvements. Whilst the latter may be beyond the power of those developing such systems, the call to involve community members at all stages of development can be addressed through diverse hiring, paid consultancy work and the like. Another avenue of community engagement is qualitative research such as the present study; the value of this form of engagement was touched upon by two interview participants, though one participant highlighted it was crucial for such work to be led by non-cisgender people. Future work should involve non-cisgender people without any familiarity with AI, through for example focus groups, to ensure a more diverse range of perspectives are captured.

Potential for good   If these issues of stereotyping, dehumanisation and sexualisation can be addressed, there is a potential for these technologies to positively represent current and yet to be imagined queer identity expressions. Interviewees felt this technology could be used to create "gender affirmative" content, and "perfectly aligned" personas, and even "[portray] queerness in ways that we haven't even thought of [which] is an exciting prospect".

Limitations

Annotation study   Our use of a small, curated set of prompts allowed for direct comparison between the models' representations of different identities. However, to investigate how these models perform *generally* when it comes to representing gender diverse identities, potentially improving the ecological validity of our annotation study, it may have been better to create a corpus of prompts through crowd sourcing or scraping image captions. This could have captured greater linguistic and cultural diversity. Our work would also benefit from extension to intersecting demographics such as disability and age.

Our annotation scheme could be extended to record "inappropriate" gendered features (for example, a transgender woman with traditionally masculine features such as facial hair). Whilst transgender women with masculine features are in no way "inappropriate", and are to be celebrated, if the

models only produce images of transgender women with stereotypically masculine features, this suggests a lack of diversity in the training data and a tendency to (re)produce stereotypes. Figure 7 in the appendix suggests this may be the case.

Survey and Interviews We surveyed non-cisgender individuals who had some familiarity with AI. While this has clear benefits, it is likely that should these tools become commercialised, the majority of those who are (negatively) impacted by their use (by the stereotyping and inaccuracy discussed in the previous section) will be those with no familiarity with the technology – the "general public". We must understand the general public's concerns and beliefs about technology in order to appropriately address these harms.

Further, by surveying those with some familiarity with AI, their proposed "solutions" may be stymied by a desire to offer solutions that seem technologically plausible. Though this has clear benefits (these solutions can become realistic medium-term goals for those developing text-to-image technologies), we may fail to uncover long-term objectives which represent how participants truly wish to be represented by such systems, current technical limitations aside. We intend to pursue a survey of the general public in future work. This will additionally allow us to compare the fears of the general public to the fears of those working AI, to understand if they align.

As noted in Section 4.2.1, our survey respondents were almost exclusively residing in the West, and were predominantly white, meaning we have failed to capture perspectives from the global south and non-white queer communities. In our interviewee selection we hoped to address this by inviting a diverse range of participants, but the interviewer's white Western background may have limited which topics participants felt comfortable discussing. Conducting the survey and interview in English will also have limited responses from non-Western individuals.

Some multiply marginalised individuals may have felt less confident in their familiarity with AI due to the Imposter Phenomenon, a reaction to "systematic bias and exclusion" know to, for example, affect women of colour in particular (Tulshyan and Burey, 2022). This may have resulted in them excluding themselves from participating where a white person with similar experience chose to respond.

Interviewees were diverse with regards to (western) gender identities, but we did not interview any transgender women, who represent a particularly vulnerable part of the community (HRC, 2022). Future work focusing on their experiences would be extremely valuable.

Finally, survey and interview participants were not compensated. Some potential respondents may have been unwilling or unable to offer free labour, again limiting the diversity of views.

Ethics Statement

Ethics approval was obtained for the annotation task, survey and interviews. In line with standard practice, we do not release the raw survey or interview data, as it contains information that may make our respondents identifiable, and we ensure that none of the direct quotes given in the paper contain any such data.

We include a brief reflexivity statement pertaining to "relevant personal and disciplinary viewpoints" (Birhane et al., 2022), and positionality statement pertaining to our "values, epistemologies, and backgrounds" (Liang et al., 2021)

The first author's interest in the representation of non-cisgender identities is driven in part by their being a member of this community. This author conducted the interviews which we hoped would address the interviewer effect – as one interview participant noted, research conducted by a cisgender interviewer would be "coloured through the lens" of their perspective (Interviewee D).

We approached this topic concerned with the potential harms these models might perpetuate through misrepresentation of the community, a concern not shared by all our survey respondents.

In addition to the limitations explored above, we identify several potential risks with this paper. Some may be offended by the images we include. We tried to mitigate this risk by including a warning in the abstract and not including images featuring genitalia. However we appreciate these images may contribute to the sexualisation and objectification of non-cisgender people, particularly if taken out of context.

Though we did not set out to generate offensive images (this would be counter to the models' intended use, for example as specified by Dayma et al. (2021)[11]), images from the full data set could

---

[11] https://huggingface.co/dalle-mini/dalle-mega

similarly offend and even be weaponised. They might accompany transphobic messages online. A data set of cisgender and non-cisgender images labeled by photorealism and presence of a clear face could feasibly be used to finetune a model to identify non-cisgender people (a concern raised by the community). As such, we make our image data set available only upon request; it is intended to measure the harm done to non-cisgender people, not contribute to it.

Acknowledgements

References

Annalisa Anzani, Louis Lindley, Giacomo Tognasso, M. Paz Galupo, and Antonio Prunas. 2021. "being talked to like i was a sex toy, like being transgender was simply for the enjoyment of someone else": Fetishization and sexualization of transgender and nonbinary individuals. *Archives of Sexual Behavior*, 50(3):897–911.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions?

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. special interest group for computing. *Information and Society (SIGCIS)*, 2.

Garfield Benjamin. 2021. What we do with data: a performative critique of data "collection". *Internet Policy Review*, 10(4).

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. (arXiv:2211.03759). ArXiv:2211.03759 [cs].

Thomas J Billard. 2019. (no) shame in the game: The influence of pornography viewing on attitudes toward transgender people. *Communication research reports*, 36(1):45–56.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 173–184, Seoul Republic of Korea. ACM.

Roland Bleiker, David Campbell, Emma Hutchison, and Xzarina Nicholson. 2013. The visual dehumanisation of refugees. *Australian Journal of Political Science*, 48(4):398–416.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv:1607.06520 [cs, stat]*. ArXiv: 1607.06520.

Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. 2022. The stable artist: Steering semantics in diffusion latent space.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2004. *Language and Identity*, page 369–394. John Wiley & Sons, Incorporated.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4568–4595, Online. Association for Computational Linguistics.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. 2021. Dall·e mini.

R DePalma and E Atkinson. 2006. The sound of silence: Talking about sexual orientation and schooling. 6(4):333–349.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 122–133, Virtual Event USA. ACM.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and Social Psychology Review*, 10(3):252–264.

HRC. 2022. An epidemic of violence 2022.

Alex Hughes. 2022. Dall-e mini: Creator explains blurred faces, going viral and the future of the project.

Johanna Kantola, Anna Elomäki, Barbara Gaweda, Cherry Miller, Petra Ahrens, and Valentine Berthet. 2022. "it's like shouting to a brick wall": Normative whiteness and racism in the european parliament. *American Political Science Review*, page 1–16.

Will Knight. 2022. Dall-e mini is the internet's favorite ai meme machine.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv:2202.11923 [cs]*. ArXiv: 2202.11923.

Calvin A. Liang, Sean A. Munson, and Julie A. Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction*, 28(2):1–47.

Rachel Minkin and Anna Brown. 2021. Rising shares of u.s. adults know someone who is transgender or goes by gender-neutral pronouns.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. 2021. Black or white but never neutral: How readers perceive identity from yellow or skin-toned emoji. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–23.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. 2020. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA. Association for Computing Machinery.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29. ArXiv:2105.02980 [cs].

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019a. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3405–3410, Hong Kong, China. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019b. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Treavian Simmons. 2018. Gender isn't a haircut: How representation of nonbinary people of color requires more than white androgyny.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*.

Ruchika Tulshyan and Jodi-Ann Burey. 2022. Stop telling women they have imposter syndrome.

Vic Valentine. 2016. *Non-binary people's experiences in the UK*.

Catherine Yeo and Alyssa Chen. 2020. Defining and evaluating fair natural language generation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 107–109, Seattle, USA. Association for Computational Linguistics.

A Annotation Task

A.1 Term Selection

We include *trans, transgender* to capture both binary and non-binary transgender identities. We

include *enby, nonbinary, gender non-conforming, genderqueer, queer* as the five most common non-binary identities (other than trans and transgender), according to the 2022 Gender Census[12] (an annual survey conducted online by a nonbinary activist). We include *two-spirit, latinx* in order to expand our focus to identities used exclusively by people of colour.

For binary identities, we combined the trans status word with *woman, man, person* and with the pronoun sets *she/her, he/him, they/them*, respectively. For nonbinary identities, we used the term *person*, with the pronouns *she/her, he/him, they/them* (it is common for nonbinary people to use both gendered and gender-neutral pronouns[13] (Dev et al., 2021)). For *two-spirit* we also combined the term with *woman, man* as we found extensive evidence online of individuals identifying as two-spirit(ed) women or men[14].

For the nonbinary identities, except *latinx* and *two-spirit* we also used the pronouns *it/it* and *xe/xem* which were the next two most common pronoun sets in the Gender Census[15]. We exclude *latinx,two-spirit* for a number of reasons: they are not well represented in the Gender Census so we felt the findings did not apply; we found no evidence of widespread use of these pronouns in either community; we felt using a potentially dehumanising pronoun such as *it* to refer to a marginalised community we did not belong to, without evidence of community use, could be harmful.

A.2 Annotation Scheme Development

We anticipate there will be less training data for non-cisgender identities and so the images will be of a poorer quality in terms of photorealism; as such we ask annotators to rate photorealism on a 4 point scale from "totally photorealistic" to "No photorealistic parts". This will also capture if images are cartoonlike: our initial experiments indicated many non-cisgender identity prompts returned cartoon-like images, which could contribute to the belief such identities are not real (Valentine, 2016; Minkin and Brown, 2021).

We ask whether an individual is present – we argue identities being depicted without a human may contribute to the "mechanistic dehumanisation" (portrayal as nonhuman and inanimate) (Haslam, 2006) of these communities.

We also ask how many humans are present. Our early experimentation on queer identities showed certain identities were often represented by the presence of two individuals.

Representations including people's faces with recognisable features are more "personal" and humanising (Bleiker et al., 2013) so we ask if a face is clearly depicted.

We anticipate that due to a lack of multiply marginalised identities being represented in the training data, most of the training data for non-cisgender identities that are not specifically associated with an ethnic minority (e.g. *latinx, two-spirit*) will be images of white individuals, and as such the models will produce mostly white figures. We ask annotators to indicate if a person of colour is present. This will be a highly subjective judgement, but we felt it was important to attempt to capture this dimension. We wished to measure whether the systems were producing clearly diverse output with regards to subject ethnicity, so we record only if the human annotator is confident that someone non-white was present. Humans tend to interpret ambiguous input as the default or norm - for example, yellow emojis are interpreted as having white ethnicity (Robertson et al., 2021).

Due to the fetishization of transgender and nonbinary individuals (Anzani et al., 2021) and the proliferation of "trans porn" online (Billard, 2019), we anticipate there will be more sexual content for prompts including non-cisgender identity terms, and so we ask annotators to indicate implied nudity. Whilst nudity is not inherently sexual, we feel greater implied nudity is a reasonable proxy for sexualisation of non-cisgender identities, and we supplement a quantitative analysis with examination of a sample of images.

Our initial experiments showed queer identities were often represented by flags instead of or in combination with people. We ask annotators to indicate if something resembling a flag is present and to describe it. Similarly, we found symbols represented often in the queer images. We also anticipate that images of non-normative identities may often be labelled, resulting in text in the image. Both of these relate to the idea of non-normative

---

[12]https://www.gendercensus.com/results/2022-worldwide/
[13]https://www.gendercensus.com/results/2022-worldwide/
[14]see for example https://www.nativeyouthsexualhealth.com/two-spirit-mentors-support-circle
[15]https://www.gendercensus.com/results/2022-worldwide/

identities being marked – that is to say, their deviation from the norm is indicated explicitly (Bucholtz and Hall, 2004). We combine these two concepts as often it is hard to distinguish computer generated letters from symbols.

One could argue a difference between cisgender and non-cisgender identity predictions as being an indicator of bias. However, we must also consider whether certain outcomes are desirable at all, even if equal e.g. should the model produce any images with implied nudity of either cisgender or non-cisgender individuals.

The annotation interface (built in the Amazon Turk sandbox) is depicted in Figure 3.

A.3 Example Output

Comparing Figures 4 and 5, there is a stark difference in the amount of nudity in response to two prompts that differ only by the word "transgender". In general we found "transgender" to elicit a lot of (partial) nudity for the Stable Diffusion model.

Also noteworthy is the absence of people of colour in both images. The model reflects the (Western) norm of whiteness.

Figure 6 demonstrates a number of "failures" – the figures rendered seem subhuman, and the model interprets Two-spirit to mean two individuals.

Figure 7 demonstrates a lack of diversity: only transgender women with features typically regarded as "masculine", such as a muscular frame or facial hair, are depicted. All the women are white. This is despite significant efforts by OpenAI to diversify DALL·E 2's output with regards to ethnicity.

B   Survey

B.1   Demographic Information

Q: What is your age? Please answer in years.
Responses: range 19-57, mode 25, mean 30.

Q1: What is your gender identity?
Options: male, female, nonbinary, genderqueer, third-gender, genderfluid, gender non-conforming, pangender, two-spirit, agender, questioning, prefer not to answer, other.

Note: A transcription error resulted in the options "male, female" in place of "man, woman" from Dev et al. (2021). Typically "male, female" are more associated with "biological sex" than "man, woman" which may have influenced respondents' answers, although the question explicitly asked about gender.

Gender	% of total responses
Male	2.9%
Female	22.9%
Nonbinary	71.4%
Genderqueer	20%
Genderfluid	8.6%
Gender non-conforming	14.3%
Agender	17.1%
Questioning	11.4%
Prefer not to answer	2.9%
Other – "trans"	2.9%
Other – "I'm also intersex"	2.9%
Other – "Woman"	2.9%

Table 2: Table of selected gender identities. Respondents could select multiple gender terms.

Sexual orientation	% of total responses
Lesbian	17.1%
Gay	8.6%
Bisexual	34.3%
Asexual	5.7%
Pansexual	17.1%
Queer	42.9%
Straight	2.9%
Prefer not to answer	2.9%
Other – "i try not to label myself "	2.9%
Other – "Bottom"	2.9%

Table 3: Table of selected sexual orientations. Respondents could select multiple terms.

Responses given in Table 2.

Q2: What is your sexual orientation?
Options: lesbian, gay, bisexual, asexual, pansexual, queer, straight, questioning, prefer not to answer, other.

Responses given in Table 3

Q3: What pronouns do you use?
Options: he/him, they/them, she/her, xe/xem, e/em, ze/hir, any pronouns, I don't use pronouns, I am questioning my pronouns, prefer not to answer, other.

Responses given in Table 4

Q4: Are you trans?
Options: yes, no, I am questioning my gender, prefer not to answer.

Responses given in Table 5

Q5: In a few words, how would you describe your ethnicity? Options: text response
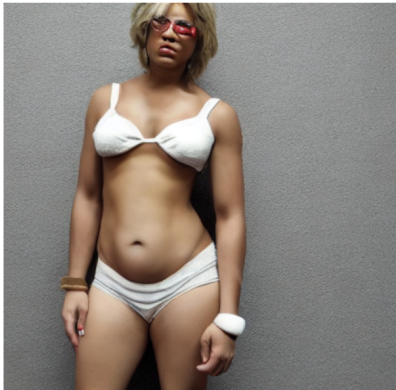
Please answer questions about the given image, then click submit!

Is the image a photorealistic picture?

[Totally photorealistic ▼]

Do you see at least one individual (photorealistic or other)?

☐ Yes

Do you see something resembling text or a symbol in the image?

☐ Yes

Do you see something resembling a flag in the image?

☐ Yes

This HIT is part of a research study. The data we collect from you is anonymous: we will receive only your worker ID together with the answers you will provide. You can find the full participant information sheet here. By continuing with this task you confirm you have read and understood the information above, and that you consent to your anonymised data being used in academic publications.

[Submit]

Please answer questions about the given image, then click submit!

Do you see at least one individual (photorealistic or other)?

☑ Yes

How many individuals are depicted in the image?

[0 ▼]

Does the image include any faces (at least half of the facial features (eyes, nose, mouth) are visible)?

☐ Yes

Is at least one of the individuals of non-white ethnicity?

☐ Yes

Is there any (implied) nudity of the crotch or torso?

☐ Yes

Do you see something resembling text or a symbol in the image?

This HIT is part of a research study. The data we collect from you is anonymous: we will receive only your worker ID together with the answers you will provide. You can find the full participant information sheet here. By continuing with this task you confirm you have read and understood the information above, and that you consent to your anonymised data being used in academic publications.

[Submit]

Figure 3: Images demonstrating the annotation interface before and after (above, below) "Do you see at least one individual" has been selected. For the commercial prompts annotators were additionally asked whether the image was relevant to the template.

Figure 4: Image generated by Stable Diffusion model in response to "Transgender men tour and enjoy the public park in summer".

Figure 5: Image generated by Stable Diffusion model in response to "Men tour and enjoy the public park in summer".

Pronoun set	% of total responses
He/him	17.1%
They/them	68.6%
She/her	34.3%
E/em	2.9%
Any pronouns	11.4%
I am questioning my pronouns	17.1%
Other - "Elle/le"	2.9%
Other - "Ey/Em"	2.9%
Other - "xey/xem"	2.9%
Other - "fae/faer"	2.9%

Table 4: Table of selected pronouns. Respondents could select multiple terms.

Response	% of total responses
Yes	85.7%
No	2.9%
I am questioning my gender	5.7%
Prefer not to answer	5.7%

Table 5: Table of responses about trans status.

Responses given in Table 7

Q8: What is/are your native language(s)?

Options: text response

The vast majority of participants (27) had English as a native language. Other native languages include German, French, and BSL.

Q9: Which country do you live in now?

Options: text response

Responses are summarised in Table 8. The vast majority of participants (34) are from Western countries, namely North America, Europe or Australia.

Q10: Briefly, how would you describe your occupation?

Options: text response

The majority of respondents (26) described themselves as explicitly white or Caucasian. Four named a European origin (none of these identified as Black, Latinx and/or Indigenous or as a person of colour); as white is the norm in Europe (Kantola et al., 2022), this suggests 30 of our 35 participants are white/Caucasian.

Q6: Are you Black, Latinx and/or Indigenous ?

Options: yes, no, prefer not to answer.

Responses given in Table 6

Q7: Are you a person of color?

Options: yes, no, prefer not to answer.

Notes: Not all respondents who identified as Black, Latinx and/or Indigenous also identified as a person of colour and vice versa.

Response	% of total responses
Yes	8.6%
No	91.4%

Table 6: Table of responses to question about identifying as Black, Latinx and/or Indigenous.

Figure 6: Image generated by `dall-e mini` in response to "Two-spirit man at a corporate event".

Figure 7: Image generated by DALL·E 2 in response to "Transgender woman at a corporate event".

Response	% of total responses
Yes	8.6%
No	91.4%

Table 7: Table of responses to question about identifying as a person of colour.

Ten respondents described themselves as students. The next most common occupation was software engineer. Other occupations include photographer, creative professional, UX designer and therapist, suggesting we were able to capture the diverse perspectives of those working outside the field but with an interest in AI.

Q11: Briefly, how would you describe your familiarity with AI?

Options: text response

The majority of respondents referenced work or education as being the source of their familiarity, though some named an interest in the topic for example as a "science magazine reader". One re-

Region	% of total responses
US	31.4%
UK	34.3%
Europe excl. UK	22.9%
Canada	5.7%
Australia	2.9%
Colombia	2.9%

Table 8: Table of responses to question about current country of residence.

spondent answered none but rated themselves as 2/5 in terms of familiarity with AI.

Q12: How would you rate your familiarity with AI?

Options: Likert scale 1-5 from "Very little knowledge" to "Expertise (I work in AI)".

Responses are summarised in Figure 8. All respondents considered themselves to have greater than "very little knowledge". The mean rating was 3.8.
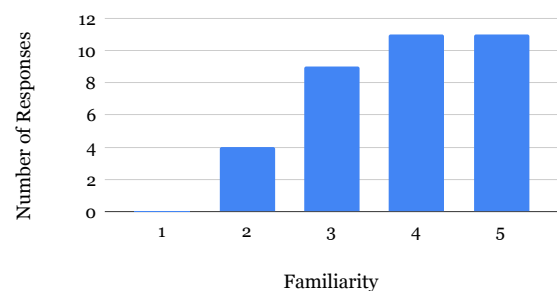

Figure 8: Count of responses for each familiarity rating.

B.2 Potential for harm.

Q13: Have you tried out one of these systems before, including during this survey?

Options: yes, no

The vast majority of respondents (28) answered yes.

Q14: Can you think of scenarios where use of

Context	% of total responses
Education	91.4%
Art/creativity	85.7%
Marketing	94.3%
Architecture/ real estate/ design	37.1%
Research	71.4%

Table 9: Table of responses to question 15 about contexts in which harm might occur to non-cisgender individuals.

Figure 9: Count of responses for each severity rating.

text-to-image models could have undesirable outcomes for non-cisgender people, due to their application in the above or other use cases?

Options: yes, no

The overwhelming majority of respondents (33) answered yes.

Q15: Please select in which of these use cases harms might occur.

Options: education, arts/creativity, marketing, architecture/ real estate/ design, research, other.

Notes: These options are derived from DALL·E 2 documentation detailing possible future commercial use of the model. A flaw in the study design meant this question was mandatory even for those who answered "no" to the previous question. Of the two participants who answered no, one wrote "none" in the "other" option and the other selected Education, but neither provided a description of a scenario (below).

Responses are summarised in Table 9. Two respondents provided "other" contexts of use – one referenced religious and political channels, and the philosophical, psychological and sociological fields, and the other wrote that they were concerned about the "reinforcement of heteronomativity in any context". The majority of respondents could imagine harm in each of the contexts except "Architecture/ real estate/ design". In particular respondents were concerned about "Marketing", "Education" and "Art/creativity" (over 3/4 of respondents felt harm might occur in these contexts).

Q16: Please select how severe you think these harms might be.

Options: Likert scale 1-5 from "No impact on lives" to "Significantly hinders lives".

Responses are summarised in Figure 9. The average rating was 3.3. Almost all participants (33) felt that the harms would have some impact on non-cisgender individuals' lives.

Q17: Please describe a specific scenario(s) where harm might occur against non-cisgender people.

Options: text response

In contrast to Dev et al. (2021) we do not ask survey participants to distinguish between representational and allocational harms, in order to reduce their work load. We label which category of harm they describe, whether it relates to how a group is represented or which services a group has access to, or both. We also identify which use cases are relevant to the harm they mention, again to reduce work load. Using a deductive-inductive approach, we also develop codes and establish themes based on the responses. One author was lead coder, developing the codebook of 17 codes. Both the lead coder and a second author applied this codebook to the responses. The coders refined the codebook through discussion, leading to a final inter-coder reliability of $\kappa = 0.74$. Themes were identified by the lead coder and discussed and finalised through discussion between all authors.

Loosely reflecting the responses to Q15, the contexts of use mentioned by respondents were education, art/creativity, marketing, and less frequently research. A high number of representational harms were identified, and very few allocational harms.

A prominent theme was the potential impact on real world behaviours and beliefs that content produced by the models might have. Frequently, respondents spoke of the output not just reflecting but *reinforcing* stereotypes and prejudices. Some felt the tools could create new beauty standards and lead to emotional harm.

Several respondents expressed concern about intentionally abusive use of these systems. They felt they might be used to create propaganda or transphobic material, or the training data needed to create a trans recognition system. Explicit references to unintentional harms were far outnumbered by

these examples.

A number of respondents explicitly referenced the role that training data played in bringing about harm, reflecting the knowledge of our respondents.

B.3 Proposed solutions

Respondents were asked to rate on a likert scale of 1-7 ("Extremely dissatisfied (I would not like to see this solution implemented)" to "Extremely satisfied (I would like to see this solution implemented)"). A rating of 4 indicates neither satisfied or dissatisfied. They were also invited to optionally respond to the question "Can you foresee any potential harms or benefits to this solution?" for each one.

Solution 1: The model generates an image based on the text (no change to current behaviour.)

Responses are given in Figure 10. Most respondents were unsatisfied with this "solution" (to change nothing), with a mode of 3 and a mean of 3.5 (both below 4). However the spread of responses indicates this is not universally disliked. Text responses in particular highlighted concerns about stereotyping,
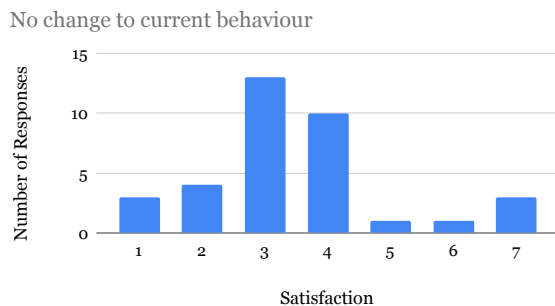

Figure 10: Count of responses for each satisfaction rating for Solution 1.

Solution 2: The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text.

Responses are given in Figure 11. This solution was the least popular, with a mode of 1 and a mean of 2. No respondents were clearly satisfied with this solution. Many respondents wrote this would lead to erasure and othering. A respondent identified it would be hard to "keep up" with queer slang, or handle ambiguous words.

A simple heuristic like ignoring minority identity terms to avoid producing stereotyped content is clearly not satisfactory to the community.

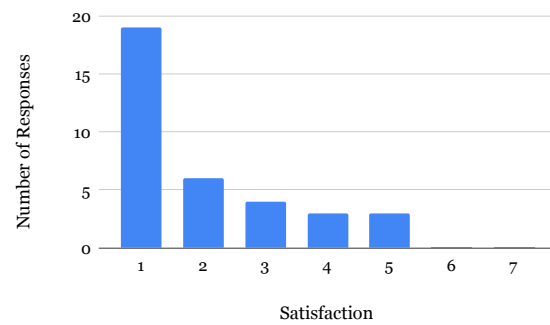**Solution 3: The model generates an image based on the text but includes a warning that the out-**

Figure 11: Count of responses for each satisfaction rating for Solution 2.

put might be offensive.

Responses are given in Figure 12. This solution was also fairly unpopular, with a mode of 2 and a mean of 3.0, although the bimodal results suggest some users would be slightly satisfied by this solution. Several respondents expressed that they felt this was not a "real" solution to the issue. Some felt strongly that appending this warning to every image suggested transness itself was offensive. However, as suggested by the second "peak", some respondents felt a warning offered an okay interim solution.
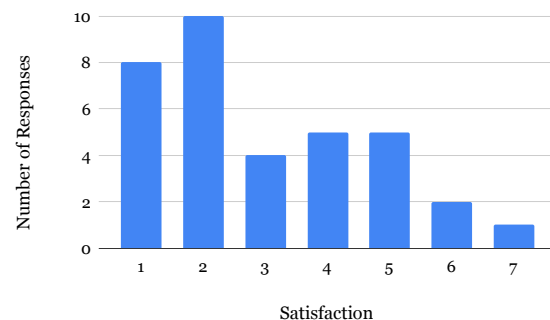

Figure 12: Count of responses for each severity rating for Solution 3.

Solution 4: The model ignores all gender identity terms in the text input and generates an image based on the rest of the text.

Responses are given in Figure 13. This solution was very unpopular, though less so than ignoring only non-cisgender identity terms, with a mode of 1 and a mean of 2.5. Some respondents expressed concern about the model "defaulting" to represent only a single gender rather than diverse results. Respondents again mentioned erasure. Several re-

spondents mentioned compromised functionality. Some felt it would be difficult to implement.
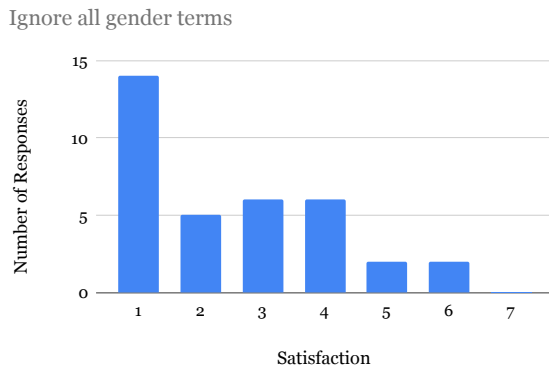
Ignore all gender terms

Figure 13: Count of responses for each satisfaction rating for Solution 4.

Solution 5: The model is trained on additional images containing non-cisgender individuals, so it better learns to generate images of non-cisgender people.

Responses are given in Figure 14. This solution was by far the most popular, with a mean of 5.3 and a mode of 7. However, as Figure 14 demonstrates, this solution is not universally popular, and in text responses respondents expressed concern about the challenge of gathering truly representative data, and the risk of reinforcing stereotypes. Some expressed concern about the risks of gathering images of marginalised people.
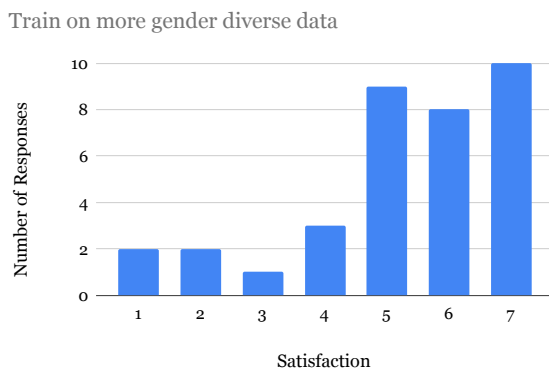
Train on more gender diverse data

Figure 14: Count of responses for each satisfaction rating for Solution 5.

Solution 6: The model effectively ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, but a flag or pin or symbol is used to indicate gender diversity.

Responses are given in Figure 15. This solution had a mode of 1 and a mean of 2.8, suggesting it

was largely unpopular (though a small number were satisfied with this solution). Some respondents expressed that this solution had potential, because it no longer required using how a person looks to capture their identity. Others felt it was a "cop out", and some were concerned about the othering or stigmatising effect of explicitly labelling queer individuals.

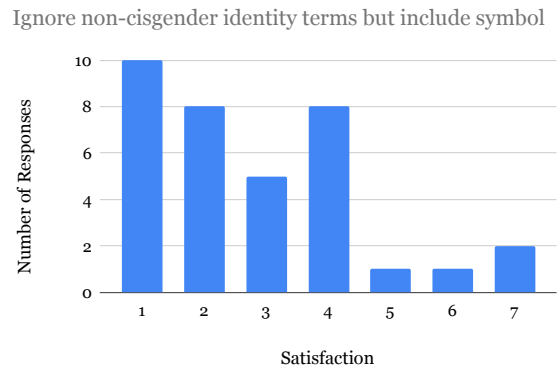Ignore non-cisgender identity terms but include symbol

Figure 15: Count of responses for each satisfaction rating for Solution 6.

Solution 7: The model ignores the non-cisgender identity terms in the text input and generates an image based on the rest of the text, with a warning that to avoid harmful misrepresentation the model ignores non-cisgender identity terms.

Responses are given in Figure 16. This solution was largely but not universally unpopular, with a mode of 1 and a mean of 2.7. Respondents expressed a preference for ignoring the terms alongside an explicit warning over simply ignoring the terms in their text responses, but many argued the same issues of erasure and compromised functionality were at play. A few saw it as a short-term solution, but many argued it was again a "cop out".
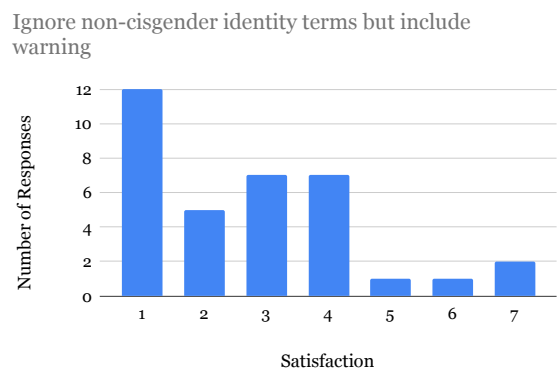
Ignore non-cisgender identity terms but include warning

Figure 16: Count of responses for each satisfaction rating for Solution 7.

Other Solutions

Respondents were then asked "Can you think of any other solutions to how models should handle non-cisgender identities? (Optional)". The majority (22) of respondents provided their thoughts. We conducted a qualitative analysis of these answers, using an inductive approach. One author developed the codebook of 22 codes using a "bottom-up" approach (driven by the data), which was then applied to the responses by a second author to establish inter-coder reliability, as a measure of code reliability. The lead coder established themes based on these codes and these themes were discussed and finalised between the authors. The major themes we established were the need for representative data; unhappiness with the proposed heuristics; the necessity of wider changes; community involvement; a desired ability to customise images.

One theme we established was the need for representative training data, echoing the most popular proposed solutions. Many respondents emphasised the need for additional data, others focused on the need to curate the training data to ensure "a diverse and representative set of images" (white, queer, nonbinary + gender nonconforming, 23).

A second theme that emerged was that of unhappiness with the proposes heuristics, with respondents seeing these as outright unsuitable or suitable only as temporary solutions.

A broad theme in the responses was the need for wider changes, encompassing both extensive changes to the model, as well as societal changes – "may require uhhh fixing society generally" (white, bisexual, genderqueer + questioning, 30). Respondents all mentioned the need to improve outcomes for other marginalised identities.

Another theme to emerge was the need for community involvement – respondents discussed the general need for non-cisgender people to be involved in the development of such models, and two suggested involving non-cisgender individuals as part of a reinforcement learning approach to improve the models' representation of the community.

The final theme represents a novel solution, which is to allow for post-hoc modification of the generated images. This would mean users could tweak the gender presentation and/or include symbols and pins to signify identity.

C   Interviews

C.1   Selecting interviewees

We selected respondents who, from their survey answers, spanned a range of gender identities, sexualities, ethnicities, occupations and countries of residence, as well as a range of attitudes towards our proposed solutions. We hoped in doing so we could ensure a diversity of opinions in our interviews over and above a random selection of interviewees.

Four of the six invited responded to our request. Our interviewees were (by their own self-reporting):

A – a white 43 year old bisexual who identifies as nonbinary (in mixed groups) and either genderfluid or agender within the queer community

B – a 33 year old pansexual nonbinary person, who identifies as "mixed race" (part Black and South American indigenous, and part Middle Eastern and white (Italian, Spanish))

C – a white Bulgarian, 30 year old bisexual genderqueer person

D – a hispanic 38 year old agender trans nonbinary person who identifies as "borderline asexual/demisexual"

C.2   Interview format

Participants were first asked a number of demographic questions about your age, gender identity, sexuality and ethnicity. Whilst we had this data already from the survey, some aspects of identity are subject to change and we wanted to ensure interview data was presented with the most appropriate descriptors.

The remainder of the interview was unstructured, with the interviewer generating questions in response to participants' answers. Participants were asked about the potential harms that could occur due to text-to-image models' handling of non-cisgender identity terms, and how participants would like such identities to be handled by these models. Participants were invited to expand on any issues raised when completing the survey.

C.3   Thematic analysis

We conducted a qualitative analysis of these answers, using an inductive approach. The coder developed an initial codebook of 41 codes using a "bottom-up" approach, then established 7 major themes based on these codes. These themes were

discussed and finalised between the authors: harmful output; being unable to use current technology; rejection of heuristics; need for community input; need for transparency and regulation; desire for authentic representation; the potential for good.

Within the theme of "harmful output", interview participants explored a range of concerns. They spoke of both unintentional harm, and deliberate weaponisation of the technology. Inaccurate representation, for example through mixing and matching of features or the enforcement of gender norms was a common topic. Participants were concerned that this misrepresentation may "set off, you know, violent stuff in the long run" (Interviewee D).

A related theme was that of being unable to use the technology in its current form: participants felt the models would not work for them easily and produce representative output as they do for cisgender people. One participant felt the technology should not be used at all.

The theme of rejecting the heuristic solutions came up in the interviews as in the survey: in particular, participants were concerned about the public associating non-cisgender identities with a offensiveness warning or maturity level label as they felt this would impact how the community is seen. Participants were also concerned about erasure due to these heuristics – " not being represented is a way to quash us right as a way to try to drive us out of existence" (Interviewee A).

As in the survey, interviewees spoke of the need for community input "at every step" (Interviewee D). They felt the greater involvement from non-cisgender and other marginalised identities there were, the more representative the output would be. One participant suggested integrating community feedback on output to capture "what that community feels is right for them" (Interviewee A). One raised the concern that these models might soon produce images "of people out of nothing without involving the people" (Interviewee C).

Another way participants suggested representation might be improved is through greater transparency and regulation. This seems particularly pertinent as several participants expressed that use of these technologies seemed inevitable. Greater transparency of training material sourcing was raised – one participant said "right now it's like we aren't acknowledge at all that humans are part of [generating training data]" (Interviewee B). Two participants were in particular concerned about the impact on artists and the need for transparency and regulation in the area of art.

A very frequent topic was a desire for authentic representation, not just of the non-cisgender community but "more representative of humanity" (Interviewee D) in general. Participants felt the training data did not reflect the reality of diversity, for example the huge global diversity of gender expressions. One participant was concerned the models would fail to represent the "different expression of gender in the global south" (Interviewee B). Respondents referenced the challenge of authentically representing communities with few members, or communities who for social, historical and technical reasons are less photographed.

Despite a number of concerns, participants did see a potential for good in these technologies. They expressed seeing both pros and cons to the technologies – "I understand that there's difficulty there, but there is also potential there" (Interviewee A); "a lot of the places where there's risks... I can see how this can be excited, exciting for another person to use" (Interviewee C). Participants saw the potential for image generation technology to be used to create "gender affirmative" output (Interviewee B), to perhaps create a persona "perfectly aligned with what you want" (Interviewee A). One participant said that "portraying queerness in ways that we haven't even thought of is an exciting prospect" (Interviewee A).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
8

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

3.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.2

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3.3, 4, 5

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3.3, 4.13, A.2, B, C.2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3.3, 4.1.1, 5, 7

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3.3, 4.1.3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
8

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.3