Responsible NLP Checklist

Paper title: *Beyond Content: How Grammatical Gender Shapes Visual Representation in Text-to-Image Models*

Authors: *Muhammed Saeed, Shaina Raza, Ashmal Vayani, Muhammad Abdul-Mageed, Ali Emami, Shady Shehata*

> How to read the checklist symbols:
>
> ☑ the authors responded 'yes'
>
> ☒ the authors responded 'no'
>
> N/A the authors indicated that the question does not apply to their work
>
> ☐ the authors did not respond to the checkbox question
>
> For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
yes, Section 5 and Appendix B.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
Yes, in Section 2 we have cited all relevant prior work, and in Table 2 we have also provide citation for all our target models

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We have mentioned the licenses for the artifacts used in our paper in Section C in the supplementary section.

☑ 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)?
Yes, in Section 4 and in Section 5.

☒ 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?
No, Our dataset consists of ground nouns words those are either grammatically masculine or feminine, and then we brought annotators to assess the gender-divergence

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Section 3

☑ 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?
Yes, in Section 3

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, Section 4.1 discusses these details and also in Limi.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Section 4.

☑ 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?
Yes, Section 5 and Table 3 provide statistics about our results, including standard errors (SE) and statistical significance indicators. Also in Appendix C, Do in Appendices D,E

N/A C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We have used APIs in our research and we have described our use in Section 4 and how generate the promts in Section 3

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

☑ 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.?
Yes, in section 3 and 4.

N/A 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)?
Section 3.5 explains that our annotators were volunteer experts who participated without payment. We describe our annotation team and their qualifications, with additional information available in Section 3 and Appendix A.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Yes, in Section 3 we are presenting how the annotation process and that participants provided informed consent for data collection.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Yes, our annotation proecess agrees with the ethics review board

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes, In section 3 and Appendix A we have described the annotators demographics/characteristics.

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
Yes, in Section 3 we have described that we have used GPT-4, Deepseek, Gemini and Claude for sampling gendered-divergence words, and also in the classification we have three models for the automatic classification process that is listed in Section 4