The authors sincerely thank the reviewers for their time and effort; your remarks will help us improve the quality of our paper, and we respond to each reviewer's queries individually below.

Reviewer 1: Thank you for the detailed thought-provoking comments which will help us improve the final work.

Overlap. Some tests reveal bias only when c-word encoding is used, and other tests reveal bias only when sent encoding is used; this suggests that a single metric does not suffice and that our method exposes biases which may otherwise be missed. We will include a detailed illustration of the overlaps (showing which tests exhibit bias in which models) in the final version.

Larger models exhibit less bias? We report two BERT model sizes (bbc_110M vs blc_370M) and the GPT model family (GPT, GPT-2_117M and GPT-2_345M). This hypothesis is supported by the number of significant positive effect sizes (Table 2), however we note that the effect sizes and significance vary in both directions across specific tests (Tables 3, 4, 5), so we do not believe we can conclude that bias "reduces" as it appears to change in nature. Because of how expensive (and inaccessible) it is to pre-train these models, we are unable to conduct a more robust study on model size and bias effect size for context dependent models. Evaluating GloVe (50d, 100d, 200d and 300d under the CBoW setup in our paper), we found no consistent trend across embedding size on effect size for either word or sent encodings.

Corpora statistics and results. We will include a test for associating M/F names and occ words in the final version.

Section 4.2: line 153. Indeed the method uses the contextual representation, so there is no pooling involved.

Reviewer 2: Thank you for the incisive comments which help us improve our discussion and interpretation of results.

Tests. We define concepts to be notions of classes like M/F and EA/AA, and define attributes to be characteristics that can be assigned to these classes, such as P/U and Career/Family. Each concept and attribute is defined with a word list (or sentence list), and thus we do not use any corpus or averaging. Regarding token-to-token similarities, specifically we calculate cosine similarities between word/sentence/contextual word encodings.

New word lists. Indeed, our word lists were constructed in prior work (Caliskan et al. [6] and May et al. [23]), which grounded the lists in the social science literature. Our contribution is in exploring them more fully across different permutations and with contextual models. E.g., for the extension of the Heilman double bind tests to race, we kept the same attribute word lists as the original tests, but replaced M/F names with EA/AA names (see also Appendix B.1-B.3).

Counting significant effects. We include Table 2 for ease of interpretation, however our main analysis is in the form of effect sizes (Tables 3-8).

Negative effect sizes. Some prior work [23] also found negative effect sizes for BERT and GPT (for sent encodings). While surprising, note that none of the instances of negative effect sizes we observed were found to be significant given the permutation test.

Intersectional results. We find that a similar proportion of our intersectional tests exhibit significant positive effects as our tests on race (25%); gender tests have a smaller proportion (12%). The experience of multiple minorities is at least as worse as their constituent minorities, but based on [12] we expected larger effect sizes in our intersectional tests. For this response, we further compared a test on M EA/F AA names with M EA/M AA and M EA/F EA names, finding that BERT exhibits larger significant effect sizes on the multiple minority case (1.57) than the others (0.68, 1.21).

Reviewer 3: Thank you for your helpful comments which will help us improving the exposition of our results.

1. In Tables 2 and 3, we do find that the c-word encoding on the non-double bind tests in general have higher effect sizes than with the sent encoding. We believe this is due to the modulating effect of pooling operations (ELMo) or the use of first (BERT) / last (GPT) word representations to obtain sent encodings.

2. Indeed, "they" is primarily used as the collective pronoun in these corpora; the takeaway observation from Table 1 is that it has more M-biased occurrences than F-biased occurrences despite being theoretically neutral.

3. We agree that terms relating to assistive devices would make sense to include. However, we determined that developing new lists was outside the scope of our expertise as it should be carefully grounded in the social science literature, and hence we only used lists developed and vetted by other scholars (e.g., [6] for ableism and age).

4. This is an interesting suggestion which we have not seen in the existing literature. It could be attained by considering negative examples (sentences which "should" be bias neutral given our understanding from the social sciences) to see if correlations are still observed. Developing such lists would be an interesting cross-disciplinary challenge.

5. There is no clear "best" method for measuring bias in any domain; indeed, it is unlikely that any single test will suffice. Rather, this work suggests that our method captures aspects of bias that other tests fail to discern.