I would like to commend the authors for their very clear and comprehensive writing of all Stage 2 sections of the manuscript. Structuring them into three units (freedom to contribute, belief about toxicity reduction, and impression that justice has been restored) greatly enhances clarity.
I also want to acknowledge the choice of the research question, which is both timely and of great practical importance - gaining insight into the best ways to respond to online toxicity. I am convinced that this study will make a valuable contribution to the ongoing debate on combating hate speech and online toxicity in general.
I have no significant comments on the manuscript; below, I offer some suggestions on how, from my perspective, the informativeness and clarity of the manuscript could be further enhanced.

1) The manuscript contains a lot of descriptive statistics spread throughout the Results section, and I have found myself looking for these descriptives in the text several times. Having all the information (e.g., M, SD, range…) for all variables and conditions contained in one table would make it clearer and much easier to find.
2) Out of curiosity, I would be interested to know how many participants failed at least one attention check. I'm wondering if a relatively small incentive may potentially lead to careless responding.
3) Within the section titled "Scale Reliability, Unidimensionality and Composites," where the authors report the results of the CFA, it would be useful to include also the results of the chi-square test, its associated df, and p-value.
4) I found it difficult to differentiate between the three conditions in Figures 6 and 7. If it were possible to increase the scale of the figures to show more units on the scale (e.g., showing more intervals like 0, 0.5, 1, 1.5, 2 instead of -2, 0, 2), it could make the differences easier to notice.
5) If a simpler explanation of the meaning of the effect sizes, particularly for H1 and H3, could be provided, it would enhance the understanding of the results. For instance, how much freer to contribute does d = 0.15 means? Perhaps offering an interpretation of unstandardized coefficients could assist in achieving this clarity.
6) In the Discussion section under the heading "Benevolent Corrections as Injunctive Norms," the authors correctly conclude that the effectiveness of this type of comments was not corroborated or that participants "neither agreed nor disagreed that benevolent corrections would lead to less toxicity from that commenter." For this reason, I find the results more relevant for discussing the role of retaliatory and benevolent going along within the Focus Theory of Normative Conduct, as suggested in the Introduction. Therefore, it seems to me that focus of the discussion could be put more on explaining the roles of retaliatory and benevolent going along behaviors that participants perceived as ineffective, rather than on the benevolent correction behavior, where the results were inconclusive (The mean for the benevolent correction condition was -0.02 on a Likert-style scale).
7) After initially reading the Main experiment discussion, I had several questions that I only found answers to after reading the General discussion. Merging both discussions (main experiment discussion and general discussion) into one coherent section could potentially provide a smoother reading experience, at least from my perspective.

2A. Whether the data are able to test the authors' proposed hypotheses (or answer the proposed research question) by passing the approved outcome-neutral criteria, such as absence of floor and ceiling effects or success of positive controls or other quality checks.

Data from manipulation checks supported the claim that the authors have successfully manipulated how benevolent and how correcting the conversations were.
To make data anonymous, I would recommend removing column „workerId" from csv data files.

2B. Whether the introduction, rationale and stated hypotheses (where applicable) are the same as the approved Stage 1 submission.

The authors have remained consistent in their framing of the study at Stage 2.

2C. Whether the authors adhered precisely to the registered study procedures.

All deviation from the original protocol are shared within the manuscript and are reasonably justified.

2D. Where applicable, whether any unregistered exploratory analyses are justified, methodologically sound, and informative.

The manuscript does not contain additional exploratory analyses.

2E. Whether the authors' conclusions are justified given the evidence.

In discussing the results, the authors strictly adhere to the data obtained.