Publication Agreement

This is a publication agreement[1] ("this agreement") regarding a written manuscript currently entitled

Weak-to-Strong Jailbreaking on Large Language Models

("the article") to be published in PMLR ("the proceedings"). The parties to this Agreement are:

Xuandong Zhao

(name of corresponding author who signs on behalf of any other authors, collectively "you") and PMLR, ("the publisher").

1. By signing this form, you warrant that you are signing on behalf of all authors of the article, and that you have the authority to act as their agent for the purpose of entering into this agreement.
2. You hereby grant a Creative Commons copyright license in the article to the general public, in particular a Creative Commons Attribution 4.0 International License, which is incorporated herein by reference and is further specified at http://creativecommons.org/licenses/by/4.0/legalcode (human readable summary at http://creativecommons.org/licenses/by/4.0).
3. You agree to require that a citation to the original publication of the article in the proceedings as well as a hyperlink to the PMLR web site linking to the original paper be included in any attribution statement satisfying the attribution requirement of the Creative Commons license of paragraph 2.
4. You retain ownership of all rights under copyright in all versions of the article, and all rights not expressly granted in this agreement.
5. To the extent that any edits made by the publisher to make the article suitable for publication in the proceedings amount to copyrightable works of authorship, the publisher hereby assigns all right, title, and interest in such edits to you. The publisher agrees to verify with you any such edits that are substantive. You agree that the license of paragraph 2 covers such edits.

[1] The language of this publication agreement is based on Stuart Shieber's model open-access journal publication agreement, version 1.2, available at http://bit.ly/1m9UsNt.

6. You further warrant that:
    1. The article is original, has not been formally published in any other peerreviewed journal or in a book or edited collection, and is not under consideration for any such publication.
    2. You are the sole author(s) of the article, and that you have a complete and unencumbered right to make the grants you make.
    3. The article does not libel anyone, invade anyone's copyright or otherwise violate any statutory or common law right of anyone, and that you have made all reasonable efforts to ensure the accuracy of any factual information contained in the article. You agree to indemnify the publisher against any claim or action alleging facts which, if true, constitute a breach of any of the foregoing warranties or other provisions of this agreement, as well as against any related damages, losses, liabilities, and expenses incurred by the publisher.
7. This is the entire agreement between you and the publisher, and it may be modified only in writing. It will be governed by the laws of the Commonwealth of Massachusetts. It will bind and benefit our respective assigns and successors in interest, including your heirs. It will terminate if the publisher does not publish, in any medium, the article within one year of the date of your signature.

I HAVE READ AND AGREE FULLY WITH THE TERMS OF THIS AGREEMENT.

- Corresponding Author:
  - Signed: Xuandong Zhao
  - Date: May 9, 2025