Responsible NLP Checklist

Paper title: *DiSec: Mitigating Backdoors in Pre-trained Language Models via Disentanglement of Adversarial Weights for Secure Fine-Tuning*
Authors: *Sunanda Das, Qinghua Li*

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?
We included the potential risks at the end paragraph of Limitations section.

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

☑ 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?
We use only publicly available benchmark datasets (SST-2, HSOL, AG News) and do not collect new user data. HSOL contains offensive/toxic language by design. We therefore avoid quoting any offensive examples in the paper. All experiments operate on tokenized inputs, and we report only aggregate metrics and mined trigger tokens (not identifiable text).

☑ 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?
We mentioned the dataset statistics in the Appendix Section B (Table 11).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We describe the experimental setup and hyperparameters (including best-found values) in Section 4 Experiments. All the experimental settings, Defense settings, Full fine-tunning settings and other settings are available here. Detailed hyperparameter tuning for DiSec are provided in the Appendix Section J Hyperparameter Tuning for DiSec

☒ 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?
We report results from a single run, and this is stated in the Full Fine-Tuning Settings paragraph under

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

the Experiments section. Due to the large number of evaluated settings (multiple models, datasets, baselines, and ablations) and common practice in previous literatures such as ONION, RECIPE, etc. we fix a single random seed and keep the same setup configuration across all experiments for consistency and fair comparison. All baselines are re-implemented under the same constraints (no downstream labels and only 20k BookCorpus sentences for statistics collection), and all fine-tuning runs follow the same training configuration.

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

N/A 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.?
No human participants or annotators were involved.

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)?
We did not recruit or pay any human participants or annotators.

N/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)?
We did not collect or curate new human-subject data, we only used existing public benchmark datasets.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We did not conduct human-subject data collection. We used existing public benchmark datasets, so no ethics board approval was required.

☑ **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?
We included the Use of Generative AI Assistance in Appendix Section L.