Responsible NLP Checklist

Paper title: *TOXIFRENCH: Benchmarking and Enhancing Language Models via CoT Fine-Tuning for French Toxicity Detection*
Authors: *Axel Delaval, Shujian Yang, Haicheng Wang, Han Qiu, Jialiang LU*

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?
Section 3.1 and Ethics Statement (Page 9) discuss the subjective nature of toxicity and the risks of dual-use..

☑ **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?
Section 3.2 and Ethics Statement (Page 9) detail the multi-step anonymization protocol and removal of PII..

☑ 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?
Section 3.2 and Section 4 provide details on the 53,622 comments, temporal distribution (20112025), and train/test splits..

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.2 and Appendix C detail the use of QLoRA, learning rates, and the SOAP optimizer..

☑ 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?
Section 4.1 (Table 3) and Section 5.3 (Table 6) report Precision, Recall, F1-score, and Accuracy with Wilson confidence intervals..

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.

☑ **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.?
Appendix F and G provide the full 010 toxicity scale and categories of implicit toxicity used for annotation..

☑ 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 3.3 specifies that human annotators were qualified native speakers compensated at 15/h..

☑ 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)?
Section 3.2 (Footnote 3) notes that data was sourced from public forums in compliance with GDPR and French intellectual property law..

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study utilized publicly available forum data and followed standard privacy/anonymization protocols rather than clinical human subject trials..

☑ **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?
Section 3.2 details the use of GPT-4o-mini for semi-automated pre-annotation and CoT reasoning generation..