Responsible NLP Checklist

Paper title: *Subject-level Inference for Realistic Text Anonymization Evaluation*
Authors: *Myeong Seok Oh, Dong-Yun Kim, Hanseok Oh, Chaean Kang, Joeun Kang, Xiaonan Wang, Hyunjung Park, Young Cheol Jung, Hansaem Kim*

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?
Ethical Considerations. PANORAMA is fully synthetic (Selvam et al., 2025). TAB uses only ECHR judgments with mandated publication and applicant consent, pre-publication de-identification, and annotation restricted to publicly inferable information (Piln et al., 2022).

☑ 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?
Ethical Considerations. PANORAMA is fully synthetic (Selvam et al., 2025). TAB uses only ECHR judgments with mandated publication and applicant consent, pre-publication de-identification, and annotation restricted to publicly inferable information (Piln et al., 2022).

☑ 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?
Ethical Considerations. PANORAMA is fully synthetic (Selvam et al., 2025). TAB uses only ECHR judgments with mandated publication and applicant consent, pre-publication de-identification, and annotation restricted to publicly inferable information (Piln et al., 2022).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 and Appendix C. Appendix C documents hardware (NVIDIA H100 80GB) and per-method hyperparameters for all four anonymization methods (Longformer threshold 0.55; DeID-GPT temp 0.05; DP-Prompt temp 1.5 / top_p 1.0; AA CoT prompt with 3 rounds). Adversarial LLM is Claude-Sonnet-4.5 (temp 0.1, selected among 11 LLMs in Appendix B); evaluator is GPT-4.1-Mini. All prompts are in Appendix G. We reproduce original-paper settings without additional hyperparameter search.

☑ 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 5.3 and Appendix E.1. We report single-run results at temperature 0.1 (near-deterministic). Robustness is assessed via three adversaries (Claude-Sonnet-4.5, GPT-4.1, Claude-Haiku-4.5) with Spearman > 0.98 for CPR and IPR across 38 configurations. IAA: TAB 93.7%, PANORAMA 95.2% (Section 3.4).

☑ **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 Appendix D. Appendix F contains the full subject-level annotation guidelines (subject rules, 15 PII categories, Hardness/Certainty rubrics, external-search policy) with a tool screenshot. Appendix D covers the entity-annotation guidelines and tool screenshot for PANORAMA span-based evaluation.

☑ 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)?
Ethical Considerations. Annotators were privacy experts and university researchers from consortium institutions of a government-funded project (PIPC & KISA, Project 2780000030), compensated through the project grant at rates adequate for professional researchers in South Korea.

☑ 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)?
Ethical Considerations and Reproducibility. TAB: ECHR judgments with mandated publication and applicant consent (MIT License). PANORAMA: fully synthetic, so real-subject consent is unnecessary (CC BY 4.0). Annotators participated voluntarily with prior notice of public release as the SPIA benchmark.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new human-subjects data was collected. Source data (TAB, PANORAMA) is already public; annotation work labels only publicly available text by voluntarily participating consortium researchers, not constituting IRB-required human-subjects research.

☑ **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?
Sections 3.3, 5.1, Appendix B, Appendix G. Claude-Sonnet-4.5 was used to pre-label 380 PANORAMA documents after comparison with 10 other LLMs (Appendix B), followed by human review and correction. In evaluation (Section 5.1), Claude-Sonnet-4.5 is the adversarial LLM and GPT-4.1-Mini is the evaluator LLM. All prompts are in Appendix G. AI assistants also supported manuscript drafting, with author verification.