Responsible NLP Checklist

Paper title: *Breaking the "Provable Security": Detecting Finite-Precision Artifacts in LLM-based Steganography via Low-Probability Vanishing*
Authors: *Wenzhao Cao, Yaofei Wang, Donghui Hu*

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 fully discuss the potential risks of this work in the Ethical Considerations section. The core discussion focuses on the dual-use nature of this security research: our findings can help defenders audit deployed steganography systems, but may also inform steganography designers of information leakage in current implementations. Accordingly, we explicitly position the proposed RRNs-HT framework as an auditing tool for controlled research and defensive evaluation, rather than an instrument for indiscriminate surveillance.

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

N/A 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?
The Movie, Twitter, and News datasets used in our experiments are widely adopted, publicly available standard academic datasets in the NLP field. All datasets have been fully de-identified by their original publishers, and contain no Personally Identifiable Information (PII) or offensive content. We did not collect or curate any new original text datasets in this work; we only used context prompts from the aforementioned public datasets for model generation, with no additional data collection or processing procedures. Thus, this item is not applicable to our work.

☑ 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 comprehensively report relevant statistics for the data used and generated in our work in Section 5.1 (Experimental Setup) and subsequent experimental results subsections: 1. We clearly specify the source and domain of the three benchmark datasets used; 2. We detail the gradient settings of generated text sequence lengths (50, 500, 2000, 5000, 8000 tokens, and 50K250K tokens under FP32 precision); 3. We clarify the unified experimental configuration and sample size for all baseline methods in comparative experiments; 4. We report averaged statistical values across the three

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.

datasets for all experimental results to ensure transparency. The proposed RRNs-HT is a non-learning statistical detection framework with no model training process, thus no train/test/dev split is involved.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We fully disclose all experimental setups and hyperparameters in Section 5.1 (Experimental Setup): 1. We specify Qwen2.5 and Llama3.2 as the backbone language models for experiments; 2. We fix the language model generation temperature at 1.0, and the top-k candidate pool size at 300 for the AC-k method; 3. We clarify the two quantization precision settings evaluated: FP16 and FP32; 4. We detail the core configuration of our homologous audit protocol (cover and stego texts are generated with the identical model family, tokenizer, floating-point precision, and global decoding policy); 5. We specify the Z-score threshold of 1.645 corresponding to the significance level =0.05 for hypothesis testing. The proposed RRNs-HT is a non-learning statistical detection framework with no training process, thus no hyperparameter search or optimal hyperparameter tuning is involved.

☑ 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 comprehensively report descriptive statistics for all experimental results in Section 5 (Experimental Evaluation), with fully transparent statistical calibers: 1. Tables 1, 3, and 4 report specific Z-score values under different configurations, with clear correspondence to experimental setups and datasets; 2. Table 2 reports the average detection accuracy of RRNs-HT and all baseline methods across the three datasets; 3. Figure 5 reports the ROC curves and corresponding AUC values to demonstrate the overall detection performance; 4. All experimental results are based on statistical analysis of multiple repeated runs, with no selective reporting of single-run results. The calculation caliber of all statistical indicators is explicitly stated in the main text.

☒ **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.?
We did not recruit or involve any human participants or annotators in this work, with no relevant experimental procedures. Thus, this item is not applicable.

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 involve any human participants or annotators in this work, with no recruitment or compensation-related procedures. Thus, this item is not applicable.

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)?
All datasets used in this work are publicly available, de-identified standard academic datasets, for which the original publishers have completed informed consent and compliance processing for the source data. We did not collect or use any raw data that can identify specific individuals, thus no procedures related to informed consent for personal data are involved.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We did not conduct any data collection activities involving human subjects in this work, with no relevant data collection protocols. Thus, this item is not applicable.

☑ **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 used AI assistants (ChatGPT) solely for grammatical polishing and phrasing improvements. No text or ideas were generated by AI.