

Responsible NLP Checklist

Paper title: *AEGIS: A Holistic Benchmark for Evaluating Forensic Analysis of AI-Generated Academic Images*

Authors: *Bo Zhang, Tzu-Yen Ma, Zichen Tang, Junpeng Ding, Zirui Wang, Yizhuo Zhao, Peilin Gao, Zijie Xi, Zixin Ding, Haiyang Sun, Haocheng Gao, Yuan Liu, Liangjia Wang, Yiling Huang, Yujie Wang, Yuyue Zhang, Ronghui Xi, Yuanze Li, Jiacheng Liu, Zhongjun Yang, Haihong E*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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 have not specifically discussed licenses or terms of use because all the assets used in this paper are open-source and comply with the respective open-source licenses.

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?
The dataset consists of academic images that do not contain personally identifying information or offensive content.

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?
Appendix A.3: Dataset Statistics

C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B.1: Experiment Environment

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 3: Experiments, Appendix B: Experiment Details

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 A.5.3: Data Quality Assurance

- 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)?

Appendix A.5.3 describes the composition of the data quality assurance team. These individuals are research collaborators involved in dataset construction and quality control, rather than externally recruited study participants, so participant recruitment and compensation do not apply.

- 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)?

This study did not involve human-subject data collection requiring participant consent.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Ethical Considerations: This study involved no human subjects or animal experimentation.

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?

Appendix E: Use of AI Assistants