## Responsible NLP Checklist

Paper title: *Bloom-Eval: A Hierarchical Evaluation Benchmark for Automatic Survey Generation Based on Bloom's Taxonomy*
Authors: *Fei Zhang, Zhe Zhao, HaiBin Wen, Tianshuo Wei, ZAIXI ZHANG, Chao Yang, Ye Wei*

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?
*The paper focuses on creating an evaluation benchmark for text generation systems and does not directly pose societal risks.*

### ☑ 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 created dataset consists entirely of published, peer-reviewed academic survey papers from 2023-2025. It contains no 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?
*Table 2 details the dataset statistics, explicitly reporting the 3,506 collected papers distributed across 60 academic venues and 14 scientific domains.*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*The implementation details, including the use of the gpt-5-mini model with a temperature of 0.0 and specific embedding models, are reported in Section 4.2.*

☑ 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?
*Detailed descriptive statistics, including F1-scores, Distributional Similarity (DS) scores, and GRADE approach scores, are reported extensively in Section 4.3.*

---

*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 G.2 describes the instructions given to the 5 independent human evaluators regarding the rating of the 30 papers across cognitive tiers.*

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)?
*Evaluators were peer researchers and graduate students who volunteered for this internal validation task.*

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)?
*The human evaluation only required participants to rate machine-generated text.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*The task involved evaluating the quality of AI-generated academic text and did not qualify as human subjects research requiring formal IRB approval.*

☑ **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?
*AI assistants were used strictly for coding assistance during the preparation of the codebase.*