

## Responsible NLP Checklist

Paper title: *DeepSynth-Eval: Objectively Evaluating Information Consolidation in Deep Survey Writing*  
Authors: *Hongzhi Zhang, Yuanze Hu, Tinghai Zhang, Jia Fu, Tao Wang, Junwei Jing, Zhaoxin Fan, Wei Bi, Ruiming Tang, Han Li, Guorui Zhou, Kun Gai*

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 did not include a dedicated discussion of potential societal or deployment risks in the current version. The paper mainly focuses on benchmark construction and evaluation. We discuss limitations of domain coverage in the Limitations section and analyze unsupported or hallucination-like outputs in Section 4.3, but we do not consider these sufficient as a standalone risk discussion.*

### 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?

*(left blank)*

- 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 3.5.*

### C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Section 4.1 and Appendix A.*

- 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 3.5 and Section 4.2.*

*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.?

*We report that six computer science experts conducted the verification process in Section 3.2, but we do not include the full annotation instructions in the paper.*

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 3.2 states that six computer science experts participated in the verification process, but the paper does not provide detailed information about recruitment or compensation.*

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)?

*(left blank)*

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

*(left blank)*

**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?

*See the Acknowledgments section. We disclose that Gemini 3 was used only for linguistic polishing and refinement, and that all AI-suggested modifications were manually reviewed and verified by the authors.*