# Responsible NLP Checklist

Paper title: *ZERA: Zero-init Instruction Evolving Refinement Agent From Zero Instructions to Structured Prompts via Principle-based Optimization*
Authors: *Seungyoun Yi, Minsoo Khang, Sungrae Park*

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?
*Yes. We discuss potential risks such as reliance on automatic metrics for summarization (e.g., ROUGE-L), potential inference latency due to longer prompts, and possible bias in internal evaluation agents under ambiguous inputs. These are detailed in the Limitations section.*

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

☑ B1. Did you cite the creators of artifacts you used?
*Yes. We used a number of scientific artifacts in our work, including benchmark datasets (e.g., GSM8K, MMLU, CNN/DailyMail), open-source LLMs (e.g., LLaMA-3.1, Qwen2.5), and prior APO systems (e.g., PromptAgent, OPRO, CriSPO). All original creators are properly cited in Sections 4 and the References.*

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*The datasets and models used (e.g., GSM8K, MMLU, LLaMA 3.1, GPT-3.5/4) are all publicly available or accessed via API under accepted terms, but we did not explicitly discuss their licenses in the paper.*

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*We used publicly released datasets and models strictly for research purposes. While our usage aligns with their intended use, we did not explicitly discuss this in the paper.*

☒ 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?
*We used only publicly available benchmark datasets that are widely used in academic research. While these datasets are assumed to be free of PII or offensive content, we did not explicitly discuss this in the paper.*

*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.*

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*We provide documentation of datasets and models used (e.g., task types and evaluation settings) in Section 4. However, we did not include detailed documentation of linguistic phenomena or demographic coverage.*

☑ 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?
*Yes. We report the number of evaluation examples (typically 500 per task) and provide details of each datasets task type and evaluation metric in Section 4. For each dataset, we use the standard test sets from public benchmarks.*

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

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*We report the names and types of models used (e.g., GPT-3.5, LLaMA-3.1-70B, etc.) in Section 4, but we did not provide details on exact parameter counts, compute budget, or infrastructure used.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Yes. We describe our iterative optimization setup (e.g., number of iterations, sample size per iteration) and scoring scheme in Section 3.23.5 and Section 4. Hyperparameter search is not applicable, as our method is training-free and does not rely on parameter tuning.*

☒ 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 report clearly whether each result is from a single run and specify the number of evaluation samples. However, we do not report error bars or multiple runs, as our framework is training-free and does not involve stochastic optimization.*

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
*While we used standard metrics such as ROUGE-L for evaluation, we did not explicitly report the implementation or parameter settings of the evaluation packages.*

☒ **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.?
*(left blank)*

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)?
*(left blank)*

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)?
*(left blank)*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*(left blank)*

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*(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?
*We used Cursor (a Copilot-based coding assistant) to assist with code completion and debugging. All experimental logic, framework design, and writing were authored by the researchers.*