Responsible NLP Checklist

Paper title: *CtrlNews: LLM-based Multi-Agent Controllable News Writing via Knowledge Gravitational Field*

Authors: *Yifei Xu, Yingjie Zong, Wang Zhonghua, Sirui Wu, Yuan Rao, Dan Zhang, Shuiguang Deng*

> 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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
All data used in this work are publicly available and open-sourced. The sources of these datasets are described in Section 4.1.

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
(left blank)

N/A 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)?
(left blank)

☑ 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 confirm that no personally identifiable information (PII) or offensive content was collected or used in our study. As described in Appendix D.4, we conducted a human evaluation to assess the quality of generated news articles. The evaluation setup did not require participants to disclose any personal details, and all responses were recorded in an aggregated form without storing individual identifiers. Thus, the collected data cannot be used to name or uniquely identify any individual.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. We provide a comprehensive description of the datasets in Section 4.1, including all data sources. In addition, Tables 22 and 23 present detailed information about the data.

☑ 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 provide a comprehensive description of the datasets in Section 4.1, including all data sources. In addition, Tables 22 and 23 present detailed information about the data.

☑ **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 did not train any models in this work but relied on existing large language models (e.g., DeepSeek-V3 and DeepSeek-R1) accessed via API for news generation. Thus, model parameters and training compute are not reported, as these are documented by the model providers. Our experiments only involve inference through API calls with minimal resources. Section 4.1 and Appendix C.2 provide the implementation details.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
This part is discussed in Appendix C.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?
This part is discussed in Appendix D.5.

N/A 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?
(left blank)

☑ **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.?
This part is discussed in Appendix D.4.

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)

☑ 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 part is discussed in Appendix D.4.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The human evaluation only involved volunteers rating the quality of generated news articles, without collecting any personal or sensitive information. Such minimal-risk evaluation is typically exempt from ethics review board approval.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
This part is discussed in Appendix D.4.

☑ **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 to help check and refine the writing content.